UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS

FACULTAD DE CIENCIAS BIOLÓGICAS

UNIDAD DE POSGRADO

Establecimiento de un índice de genes en Ipomoea batatas (L.) Lam. usando secuenciamiento 454 a partir de bibliotecas de cDNA y desarrollo de marcadores microsatélites

TESIS

para optar al grado académico de Magister en Biología Molecular

AUTOR

Tincopa Marca, Luz Rosalina

Lima-Perú

2010

AGRADECIMIENTOS

- Este trabajo se realizó en los laboratorios del Centro Internacional de la Papa (CIP) y con el finaciamiento completo tanto del CIP como del *The Generation Challenge Programme (GCP)*.

- Al Consejo Nacional de Ciencia y Tecnología (CONCYTEC) por haberme otorgado una beca de estudios de maestría, dinero que me ha servido para asistir a cursos y conferencias que de alguna u otra forma ha contribuido a una mejor preparación académica en el desarrollo de mi tesis.

- A mi asesor, Roland Schafleitner, por permitirme desarrollar este tema de tesis y guiarme en el proceso del mismo.

- A Omar Palomino por el soporte técnico durante el proceso del trabajo de esta tesis, de quién he aprendido mucho sobre informática en todo este tiempo. A Luis Rojas, Jack Hou, Maria Vargas, Raymundo Gutierrez, Julio Solis, Ronald Robles, Rocio Alagon, Carlos Rivera, Cynthia Quispe, Ji Young Kim y Foo Cheung, quienes me han ayudado en algún proceso del desarrollo de la tesis.

- A mi jurado de tesis conformado por los profesores: Patricia Woll, Ruth García, Giovanna Sotil, Rina Ramirez y Pablo Ramirez, quienes me han ayudado diligentemente en las correcciones de la tesis.

- A mi familia que siempre ha estado para apoyarme y en especial a mi madre Luz Marca y mi padre Jorge Tincopa por el amor y la confianza que siempre me brindan, a mis hermanos Rosario Mochcco, Kervin, Jhon, Cristhian, Yuri, Rosa y Milagros Avalos y a mi tio Janio Avalos, por siempre apoyarme.

- A Rosario Herrera quien me dio la oportunidad de ingresar al CIP, a Carlos Merino por ser mi guía los primero días en el CIP, a Amelie Gaudin quien ha contribuido en mi formación científica y a toda la gente del CIP con los que he compartido

experiencias fuera y dentro del laboratorio antes, durante y después del proceso de realizar la tesis: Lina Bernaola, Rebeca Schibli, Percy Rojas, Cithya Zorrilla, Dora Quispe, Karin Cruzado, Gabriela Lajo, Kathy Prentice, José Tovar, Milagros Ormachea, Milton Untiveros, Diógenes Cerna, Julian Soto, Jorge Biondi, Martin Ramos, Macarena Barra, Julio Frisancho, Bruno Lértora, Lorena Guardia, Diana Díaz, Ada Sumi, Juan Montenegro, Kelvin Huamaní, Federico Díaz, Jeanet Orbegoso, Romina Reaño, Mariela Aponte, Luciano y Max Fernandez, Luis Gutierrez, Joel Reyes, José Rodriguez, Alberto Bendezú, y todos los que me he olvidado en esta gran lista de gente que quiero mucho y que ha hecho que mi estadia en el CIP sea agradable, divertida e inolvidable.

"I am among those who think that science has great beauty. A scientist in his laboratory is not only a technician: he is also a child placed before natural phenomena which impress him like a fairy tale"

Marie Curie

(1867-1934)

A mí maestro.

CONTENIDO **Pág**

LISTA DE TABLAS **Pág.**

LISTA DE FIGURAS **Pág.**

LISTA DE ABREVIATURAS

E-value	Expectation Value
BLAST.	Basic Local Alignment Search Tool
EST	Expressed Sequence Tags
Contig	Uniones de sobrelapamiento de EST
Singleton	EST sin ensamblar
DNA	Ácido desoxirribonucleico
EC number	Enzyme Classification number
GI number	Unique Identifier in the GenBank database
GO number	Gene Ontology number
mRNA	RNA mensajero
NCBI	National Center for Biotechnology Information
NCBI nr	NCBI non-redundant protein sequence database
ORF	Open reading frame (marco abierto de lectura)
RNA	Ácido ribonucleico
UniProt	Universal Protein Resource
PCR	Reacción en cadena de la polimerasa
cDNA	Complementary DNA
MMP	Mínimun match percent (porcentaje mínimo de identidad en los sobrelapamientos)
dNTP	Deoxyribonucleotide
GO terms	Término del *Gene ontology*
Mers	Secuencias únicas nucleotídicas, que se encuentran en el sobrelapamiento de los *reads*
Reads	EST
GenBank	Base de datos de secuencias genéticas del *National Institutes of Health de Estados Unidos*.

RESUMEN

Un índice de genes de camote *Ipomoea batatas* (L.) Lam. ha sido establecido basado en el pirosecuenciamiento. Dos colecciones normalizadas de cDNA de tallos y hojas de camote cultivar Tanzania que habían sido expuestos a sequía, dieron 524,209 *reads*. Después del establecimiento de los parámetros óptimos de ensamblaje, estos *reads* fueron ensamblados junto con 22,094 EST disponibles públicamente en el GenBank, dando como resultado del ensamblaje 31,685 *contigs* y 34,733 *singletons*. Las comparaciones usando Blastx en la base de datos del UniRef100 permitió la anotación de 23,957 *contigs* y 15,342 *singletons*, resultando en 24,657 genes únicos. Además, 27,119 secuencias no tienen ninguna coincidencia con las secuencias proteicas del Uniref100. La anotación de 24,763 secuencias incluye la atribución de *GO terms*. Adicionalmente, se han encontrado 293 genes involucrados en la respuesta a estrés hídrico. Basado en este índice de genes, se ha identificado 195 marcadores microsatélites que fueron amplificados exitosamente y evaluados en un grupo de seis hexaploides de *Ipomoea batatas* y dos diploides de *Ipomoea trífida.*

Palabras claves: anotación de genes, ensamblaje de transcriptoma, *Ipomoea batatas*, *Ipomoea trífida*, marcadores microsatélites, pirosecuenciamiento 454, recursos genéticos.

ABSTRACT

A sweetpotato *Ipomoea batatas* (L.) Lam. gene index was established based on pyrosequencing. Two normalized cDNA collections from stems and leaves were produced from drought stressed sweetpotato cultivar *Tanzania* and yielded 524,209 pyrosequencing *reads*. After establishment of the optimal assembly parameters, these *reads* were assembled together with 22,094 publically available expressed sequence tags into 31,685 sets of overlapping DNA segments and 34,733 unassembled sequences. Blastx comparisons with the UniRef100 database allowed annotation of 23,957 *contigs* and 15,342 *singletons* resulting in 24,657 putatively unique genes. Further 27,119 sequences had no match to protein sequences of UniRef100 database. Annotation of 24,763 sequences included the attribution of gene ontology terms to as many sequences as possible. In adition it was found 293 genes involved in water response. Based on this gene index, we have identified 195 gene-based microsatellite markers that could be successfully amplified and scored in a test panel of six hexaploid *I. batatas* and 2 diploid *I. trifida* genotypes.

Key words: *Ipomoea batatas, Ipomoea trífida*, gene annotation, genetic resources, microsatellite markers, 454 pyrosequencing, transcriptome assembly.

1 INTRODUCCIÓN

El camote [*Ipomoea batatas* (Linnaeus, 1753) Lamark] es el séptimo cultivo más importante en términos de producción a nivel mundial. Este cultivo se adapta muy bien a diferentes medio ambientes y es relativamente tolerante a sequía y calor. El camote es principalmente cultivado en los países en desarrollo, el cual representa más del 95% de la producción mundial. El 80% de la producción crece en Asia, el 15% en África y el 5% en el resto del mundo. En Sudamérica el camote es el tercer cultivo de raíces más importante después de la yuca y la papa, con una producción de 1.27 millones de toneladas anual y con un rendimiento de 17.2 t/ha. China es el mayor productor de camote en el mundo, con una producción anual de 85.2 millones de toneladas y un rendimiento de 23.1 t/ha. A pesar de la resistencia por la que se caracteriza este cultivo, la producción de camote en África es muy baja, 14 millones de toneladas anual y un rendimiento de 4.2 t/ha (FAOSTAT, 2008). Este bajo rendimiento se debe a la falta de calidad de las semillas, problemas con insectos, nemátodos y virus (Loebestein y Thottappilly, 2009).

La importancia de este cultivo radica en su valor nutritivo. Las raíces poseen un alto contenido de carbohidratos y algunos genotipos contienen cantidades significativas de hierro y zinc, y en especial el camote de color amarillo o naranja posee un alto contenido de betacaroteno, precursor de la vitamina A (Bovell-Benjamin, 2007). Esta propiedad nutritiva lo hace idóneo para el uso como fuente de nutrición en países donde existe una deficiencia de vitamina A.

Centro América es una de las regiones que posee una gran biodiversidad de camote en el mundo seguido por Sudamérica. En el Perú existe una gran biodiversidad de camote y el Centro Internacional de la papa tiene un banco de germoplasma de camote con 5,960 accesiones, que posee tanto variedades del Perú como de otras regiones del mundo (CIP sweetpotato database at http://germplasmdb.cip.cgiar.org/biomart/martview/79cf53d5d8b717e9f0da704e48345792).

Debido a sus propiedades agronómicas y nutricionales, preservar e incrementar el rendimiento de este cultivo es una manera excelente de combatir la desnutrición en los países en desarrollo. Sin embargo, debido a la complejidad genética de esta planta

hexaploide, el mejoramiento de variedades bien adaptadas, resistentes contra enfermedades a plagas y tolerantes a estrés abiótico, es difícil de obtener y lleva mucho tiempo. La aplicación de programas de mejoramiento genético se ve obstaculizada por la falta de recursos genéticos y genómicos para este cultivo.

Actualmente la información genómica de *Ipomoea batatas* es muy limitada, pero hoy en día, con las últimas tecnologías de secuenciamiento y análisis bioinformáticos nos permiten obtener información de la secuencia de genes a bajo costo. Esta información provee conocimiento acerca del contenido genético de un organismo y al mismo tiempo permite el diseño de marcadores de selección que pueden ser usados como herramientas en mejoramiento genético. El presente trabajo está dirigido a la generación de información genética y genómica de camote útil como herramientas para los programas de mejoramiento genético de este cultivo en el futuro. Este objetivo está de acuerdo a los lineamentos del Centro Internacional de la Papa (CIP) que es el de contribuir a la salud humana y reducir la pobreza.

En este sentido, los objetivos de este trabajo fueron los siguientes:

Generales:

- Establecer un índice de genes en *Ipomoea batatas* (L.) Lam.
- Identificar y desarrollar marcadores microsatélites.

Específicos:

1. Producir bibliotecas de cDNA de hojas y tallos de *Ipomoea batatas* (L.) Lam.
2. Ensamblar las secuencias generadas por 454 y EST del Genbank de *Ipomoea batatas* (L.) Lam.
3. Anotar los genes de *Ipomoea batatas* (L.) Lam.
4. Identificar microsatélites y diseñar iniciadores a partir del índice de genes.
5. Amplificar los marcadores microsatélites.

2 ANTECEDENTES

2.1 El camote

El camote es una dicotiledónea rastrera perenne que pertenece a la familia Convolvulaceae. El camote posee hojas en forma de corazón o palmeteadas lobuladas y posee flores simpétalas. Posee raíces ricas en almidón; el color de la cáscara varía entre rojo, morado y blanco y el color de la pulpa varía entre blanco, amarillo naranja y morado. Las hojas jóvenes y los brotes algunas veces son consumidos como verduras. El camote es una planta hexaploide altamente heterocigótica, se propaga vegetativamente y por reproducción sexual, el número de cromosoma base es de n=15, todavía no se ha determinado si tiene un origen autohexaploide o alohexaploide (Sweetpotato, CGIAR).

2.2 Origen del camote

El camote es nativo de la parte tropical de Sudamérica y fue domesticado por lo menos 5,000 años atrás (Sweetpotato, CGIAR). Austin en 1988 propuso que el centro de origen de *Ipomoea batatas* está entre la Península de Yucatán en México y la boca del río Orinoco en Venezuela, y que de ahí se difundió hacia el Caribe y Sudamérica hace 2500 años antes de Cristo. Zhang *et al*. (1998) proporcionan una fuerte evidencia sobre el origen del camote a la misma zona geográfica propuesta por Austin, como centro primario de diversidad, y como centro secundario al Perú y el Ecuador, debido a que existe una diversidad molecular más baja. El camote también crece en Polinesia donde es conocido como "*kumara*" y se cree que existió antes de las exploraciones europeas. Se ha encontrado restos de camote en las islas Cook que datan de 1000 años antes de cristo. Se piensa que los polinesios viajaron a Sudamérica y de regreso llevaron y diseminaron el camote a través de la Polinesia, Hawai y Nueva Zelandia (Clarke, 2010).

2.3 Cultivo del camote

El camote crece en las regiones tropicales, subtropicales y en regiones cálidas. La planta no tolera la helada. El camote es propagado a través de cortes de esquejes de 20 a 30 cm de largo que directamente son plantadas en el suelo, dependiendo del cultivar y de las condiciones, las raíces pueden obtenerse entre los dos a nueve meses (Woolfe, 1992). Este cultivo es sensible a sequía en la fase inicial de tuberización, entre los 50-60 días después de haberse plantado pero en estado de desarrollo tardío, muchos genotipos de camote son altamente tolerantes a sequía. El camote tampoco es tolerante a las inundaciones, ya que esto puede causar la necrosis de las raíces y reducir el crecimiento de las raíces si la aireación es pobre (Ahn, 1993).

2.4 Producción y rendimiento del camote

La producción global de camote ha sido cerca de 110 millones de toneladas en el 2008 y la producción por continentes fue de 92.49, 14, 2.85, 0.7 y 0.067 millones de toneladas en Asia, África, América, Oceanía y Europa respectivamente (Figura 1). Claramente Asia es el mayor productor, y sólo China tuvo una producción de 85.2 millones con un rendimiento de 23.1 toneladas por hectárea, en contraste el rendimiento de camote está alrededor de 4.2 toneladas por hectarea en África (FAOSTAT, 2008). Existen varios factores responsables para el bajo rendimiento en África, primero la disponibilidad de variedades de camote mejorado está limitada en este continente, por lo tanto, los agricultores sólo dependen de las variedades con bajo rendimiento. Los rendimientos se reducen aún más por las enfermedades y plagas como el virus del camote que pueden causar hasta un 98% de la reducción del rendimiento (Loebestein y Thottappilly, 2009).

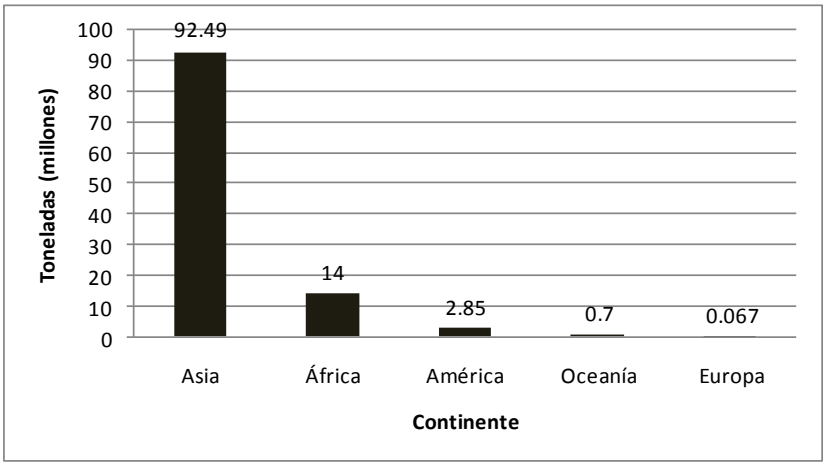

Figura 1. Producción de camote por continentes en el 2008 en millones de toneladas (FAOSTAT, 2008).

2.5 Nutrición

El camote contiene muchos nutrientes, dentro de los cuales incluyen a las proteínas, carbohidratos, minerales (calcio, hierro, potasio) carotenoides, fibra dietética, vitamina C, vitamina B6, muy poca grasa y sodio (Bovell-Benjamin, 2007). Las variedades rosadas y amarillas tienen alto contenido en betacaroteno, precursor de la vitamina A. Las hojas de camotes son ricas fuentes de minerales y sustancias polifenólicas que tienen varias funciones fisiológicas en la planta y proporcionan beneficios para la salud de los consumidores, tales como la actividad antioxidante, antimutagénica, anticancerígena, antidiabética y antibacterial (Islam, 2006).

Los carotenoides, particularmente el betacaroteno, un precursor de la vitamina A, juega un rol importante en la salud. La deficiencia de vitamina A todavía es uno de los mayores problemas de salud que aqueja a los países en desarrollo. A nivel mundial, 140 millones de niños menores de 5 años tienen una falta extrema de vitamina A (Van Jaarsfeld *et al*., 2005, Mason *et al*., 2001). Por ejemplo, África oriental y meridional tiene la más alta prevalencia de niños en edad pre escolar con riesgo de deficiencia de vitamina A, donde más de 3 millones de niños sufren de ceguera a causa de la falta de esa vitamina. La deficiencia de vitamina A es también una de las causas importantes de muerte temprana en niños y un factor de mayor riesgo para las mujeres embarazadas en África. Los camotes con alto contenido de betacaroteno tienen un gran potencial de mejorar la disponibilidad de

pro-vitamina A para la población de bajos recursos en estas regiones. El camote de pulpa naranja provee cerca de 350 RAE (equivalentes de actividad de retinol) por cada 100g de materia fresca, los cuales contienen 4,200µg de betacaroteno (Low *et al*., 1997, 2007), que puede suplir la cantidad diaria de vitamina A que requiere un niño. Además, en la actualidad existen variedades de camote de pulpa naranja que contienen de 20 a 30 veces más betacaroteno que el arroz dorado (Ye *et al*., 2000). El fomento de cultivares de camote con alto contenido de provitamina A es muy importante para combatir la malnutrición en muchas regiones de África. Sin embargo, para el desarrollo de variedades adaptadas se necesita inversión en el mejoramiento de cultivos. Un componente de esta inversión son las herramientas de mejoramiento tales como la información de secuencias génicas y marcadores moleculares.

2.6 Recursos genómicos y genéticos del camote

El CIP posee 5,960 accesiones de camotes y especies familiares silvestres de camote, que representa la biodiversidad de esta especie. Un set compuesto de genotipos consistente de 472 accesiones, que representa la diversidad de este cultivo, ha sido identificado y está disponible para la distribución de los mejoradores en el Centro Internacional de la Papa (http://gcpcr.grinfo.net/files/cr_files/gcpcr_file832.xls). Con la finalidad de usar el germoplasma de camotes en programas de mejoramiento genético es necesario desarrollar herramientas moleculares tales como secuencias genómicas, marcadores moleculares y mapas genéticos. La información sobre un marcador asociado a un carácter tiene el potencial de hacer más eficiente y más directa la selección. El acceso a secuencias de genes de camote podría facilitar la aplicación de procedimientos de genómica funcional investigando vías metabólicas y aclarar mecanismos de resistencia contra estrés biótico y abiótico. Sin embargo, la información genómica para camote es escasa, actualmente hay 22,371 ESTs depositados en el Genbank, 1,837 secuencias nucleotídicas, 641 secuencias proteícas, 6 estructuras tridimensionales proteícas (NCBI Mayo 2010), un mapa genético basado en marcadores AFLP (Kriegner *et al*., 2003, Cervantes-Flores *et al*., 2008) y se han reportado marcadores microsatélites (Jarret y Bowen, 1994; Buteler *et al*., 1999; Zhang *et al*., 2000; Hu *et al*., 2004; Arizio *et al*., 2008; Veasey *et al*., 2008). También se han usado algunos marcadores RAPD y AFLP para estudios de diversidad genética en variedades chinas de camote (He *et al*., 2006).

2.7 Síntesis y normalización de la biblioteca de cDNA

La biblioteca de cDNA es la vía más eficiente para obtener secuencias que representan genes expresados. Esta estrategia es el punto de inicio para caracterizar la variación genética funcional en organismos. Las bibliotecas de cDNA son producidas por la transcripción reversa de mRNA y así éstas representan copias de DNA de los transcriptos. Estos cDNAs luego pueden ser clonados para más análisis o pueden ser directamente secuenciados. Una de las técnicas utilizadas para producir bibliotecas de cDNA es *SMART*^{TM} *Technology*. Este método utiliza la actividad de "cambio de plantilla" de la transcriptasa reversa del *Moloney murine leukemia virus* (MMLV) para sintetizar y amplificar el cDNA de hebra simple en un solo paso, siguiendo la transcripción reversa, tres ciclos de PCR son realizados usando un primer oligo (dT) modificado y un primer de anclaje para enriquecer la población de cDNA para secuencias de tamaño completo. Este método presenta ventajas como obtener un gran rendimiento de cDNAs de tamaño completo, minimiza la pérdida de las secuencias en el extremo 5' y se puede partir de pequeñas cantidades de RNA para la síntesis de cDNA además requiere menos tiempo y pasos en comparación a otros métodos (Zhu *et al*., 2001).

2.8 Construcción de la biblioteca de cDNA usando *SMART*^{TM} *Technology*

La síntesis del cDNA de hebra simple se realiza en la presencia de dos oligonucleótidos "*CDS primer"* que contiene un sitio de restricción para la enzima *Sfi*IB y el "*TS-oligonucleotide"* que contiene un sitio de restricción para la enzima *Sfi*IA. Cuando la transcriptasa reversa llega al extremo 5' del mRNA, la actividad transferasa que posee esta enzima añade nucleótidos adicionales (predominantemente dCTP) que no están codificados por la plantilla original, en la hebra de cDNA que está siendo sintetizada. El "*TS-oligonucleotide"* contiene tres nucleótidos consecutivos de guanina en el extremo 3' del "*TS-oligonucleotide"* que sirve como una segunda plantilla para la transcriptasa reversa. Cuando la transcriptasa reversa llega al extremo 5'del mRNA, la interacción complementaria de los tres nucleótidos de guanina en el extremo 3' del "*TS-oligonucleotide"* y la secuencia extendida rica en citosina del cDNA promueve el cambio

de plantilla. La transcriptasa reversa transcribe a continuación el oligonucleótido, añadiendo la secuencia nucleotídica SfiIA hasta el final en el cDNA de hebra simple que está siendo sintetizada. Luego el componente RNA del híbrido cDNA-RNA es degradado por el tratamiento con NaOH. A continuación el cDNA de doble hebra es generado con tres ciclos de PCR catalizado por la mezcla de reacción que contiene polimerasa para fragmentos largos. Esta reacción usa un iniciador que es complementario a la secuencia *Sfi*IA y el "*CDS primer"*que contiene la secuencia *Sfi*IB. Adicionalmente si uno quiere realizar el clonamiento de la secuencia, se realiza la digestión del cDNA de doble hebra con SfiI que genera dos diferentes extremos cohesivos, SfiIA y SfiIB, en el extremo 5' y 3' respectivamente. Debido a que los sitios *SfiI* son extremadamente raros en el DNA de mamíferos, casi todos los fragmentos de cDNA se mantienen intactos, eliminando la necesidad de metilación del cDNA. Después de la digestión del cDNA, este es clonado direccionalmente en λTrip1Ex2, un vector fagémido que subclona mediante el sitema *cre-lox* (Zhu *et al*., 2001).

Existen dos principales tipos de bibliotecas de cDNA: normalizada y no normalizada. En la biblioteca no-normalizada, el número de cDNA deriva de un gen específico generalmente correlacionado con su nivel de expresión. La normalización tiene como objetivo reducir la diferencia en el número de copias de los cDNAs derivados de los genes con alta y baja expresión. Así, mientras la biblioteca no-normalizada de cDNA nos da una visión del patrón de expresión de genes, las bibliotecas normalizadas nos dan un mejor punto de vista sobre el contenido total del transcriptoma de un tejido. Como la normalización dirige a una reducción de los genes altamente expresados, entonces se hace fácil identificar los genes con baja expresión (Zhulidov *et al*., 2004).

Un método generalizado para producir una biblioteca de cDNA normalizada es mediante el sistema de *duplex-specific nuclease* (DSN). La normalización DSN involucra la deshibridización y reasociación del cDNA, la degradación de las fracciones de DNA doble hebra formado por los transcriptos abundantes y la amplificación de PCR de la fracción de DNA de hebra simple. La etapa más importante de este método es la degradación de la fracción de doble hebra formado durante la reasociación del cDNA usando el DSN de cangrejo "*Kamchatka"*. Esta enzima termoestable presenta una fuerte preferencia por las dobles hebras DNA-DNA o DNA-RNA en comparación por la simple hebra de DNA o RNA, independientemente de la longitud de la secuencia (Zhulidov *et al*., 2004). Este

método está disponible en *kits* para usuarios o como un servicio ofrecido por compañías biotecnológicas.

2.9 Secuenciamiento 454

El secuenciamiento 454 es una de las últimas tecnologías de pirosecuenciamiento desarrollado por Margulies *et al*. (2005). El sistema de secuenciamiento 454 puede analizar diversos tipos de muestras: DNA genómico, producto de PCR, BACs y cDNA. El DNA es cortado en pequeños fragmentos usando un proceso denominado nebulización. Para las muestras pequeñas como los RNA no codificante o amplicones de PCR, la fragmentación no es requerida. Se agrega pequeños adaptadores (A y B) a cada fragmento, los cuales son usados en el proceso de purificación, amplificación y secuenciamiento (Figura 2). Luego, los fragmentos de hebra simple (sstDNA) unidos a los adaptadores componen la biblioteca de la muestra usado para los siguientes pasos del secuenciamiento. Durante el siguiente procedimiento de emulsión de PCR, la biblioteca de sstDNA es primero mezclada con *beads* de *sepharosa* que contienen los oligonucleótidos complementarios a los adaptadores A y B (Figura 3). La relación de la concentración de los *beads* y el DNA favorece la generación de la asociación de *DNA-beads* que carga un único fragmento de DNA. Luego, la biblioteca de los *beads* es emulsificado con reactivos de amplificación en una mezcla de agua y aceite. Cada *bead* es luego capturado dentro de su propio microreactor donde la amplificación clonal del fragmento de hebra simple ocurre (Figura 4). Cada *bead* unido al DNA es colocado en un pocillo en el PicoTiterPlate, un chip de fibra óptica. El tamaño del pocillo del *PicoTiterPlate* sólo permite ingresar un *bead* por cada pocillo. La biblioteca sstDNA luego es agregada a una mezcla de incubación conteniendo DNA polimerasa y una cubierta con las enzimas en *beads* conteniendo sulfurilasa y luciferasa sobre el 454 PicoTiterPlate, luego el PicoTiterPlate es colocado dentro del GS FLX para el secuenciamiento. El sistema de flujo proporciona los reactivos de secuenciamiento (conteniendo el tampón y nucleótidos) a través de los pocillos de la placa. Los nucleótidos son agregados en forma secuencial en un orden fijo a través de PicoTiterPlate durante el proceso de secuenciamiento. Durante el flujo de nucleótidos, cientos de miles de *beads* cada uno cargando millones de copias de moléculas únicas de DNA de hebra simple son secuenciados en paralelo. Si un nucleótido complementario a la hebra plantilla fluye dentro de un pocillo, la polimerasa extiende la hebra de DNA existente por la adición de

nucleótidos (Figura 5). La adición de uno o más nucleótidos resulta en una reacción que genera una señal de luz que es grabada por una cámara con un dispositivo de cargas eléctricas interconectadas (*charge-coupled device*-CCD) en el instrumento. En un rango limitado, la intensidad de la señal es proporcional al número de nucleótidos incorporados en un flujo (Figura 6). Esta técnica se basa en la secuenciación por síntesis y se llama pirosecuenciación (Ronaghi *et al*., 1998).

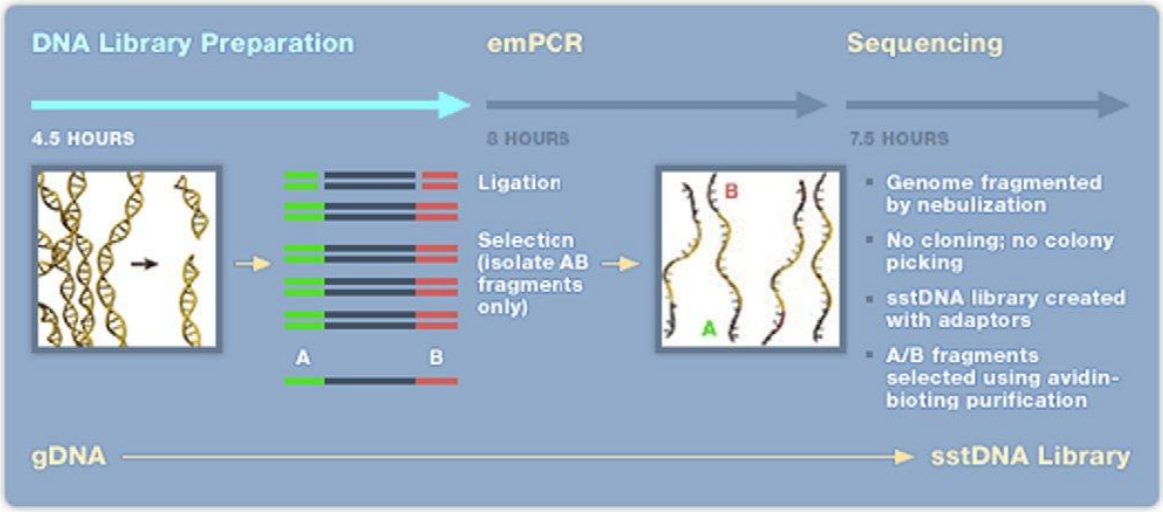

Figura 2. Generación de fragmentos de cadena sencilla con adaptadores A y B (Droege y Hill, 2008).

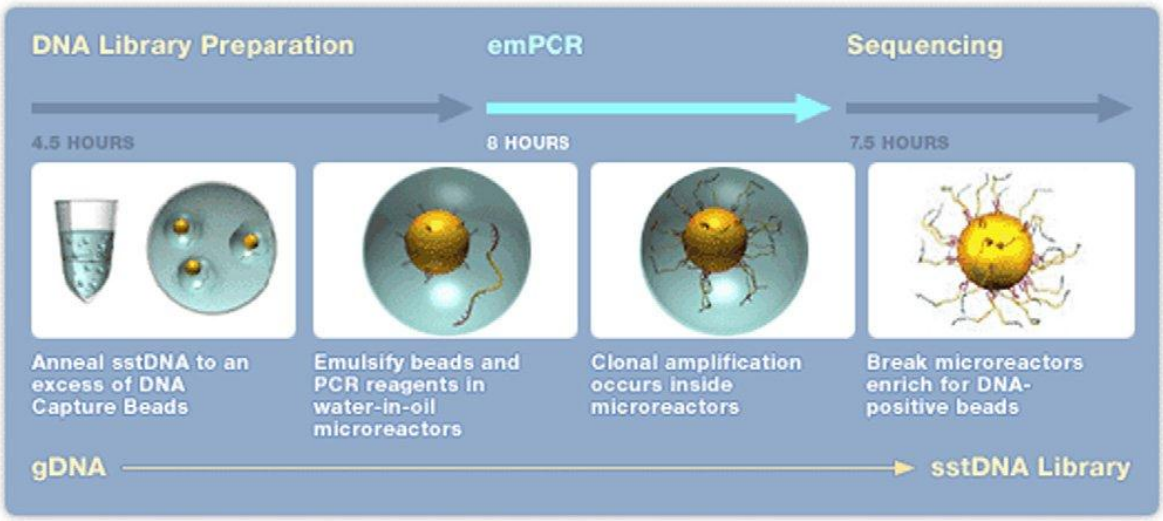

Figura 3. Amplificación clonal de sstDNA sobre los beads durante la emulsión de PCR (tomado de Droege y Hill, 2008)

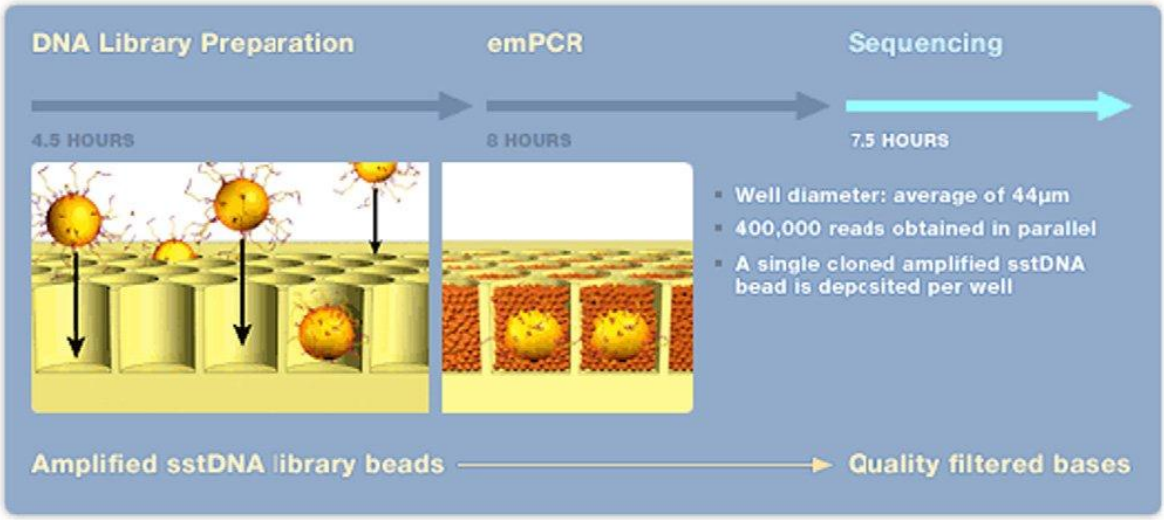

Figura 4. Amplificación clonal del fragmento de hebra simple ocurre (tomado de Droege y Hill, 2008).

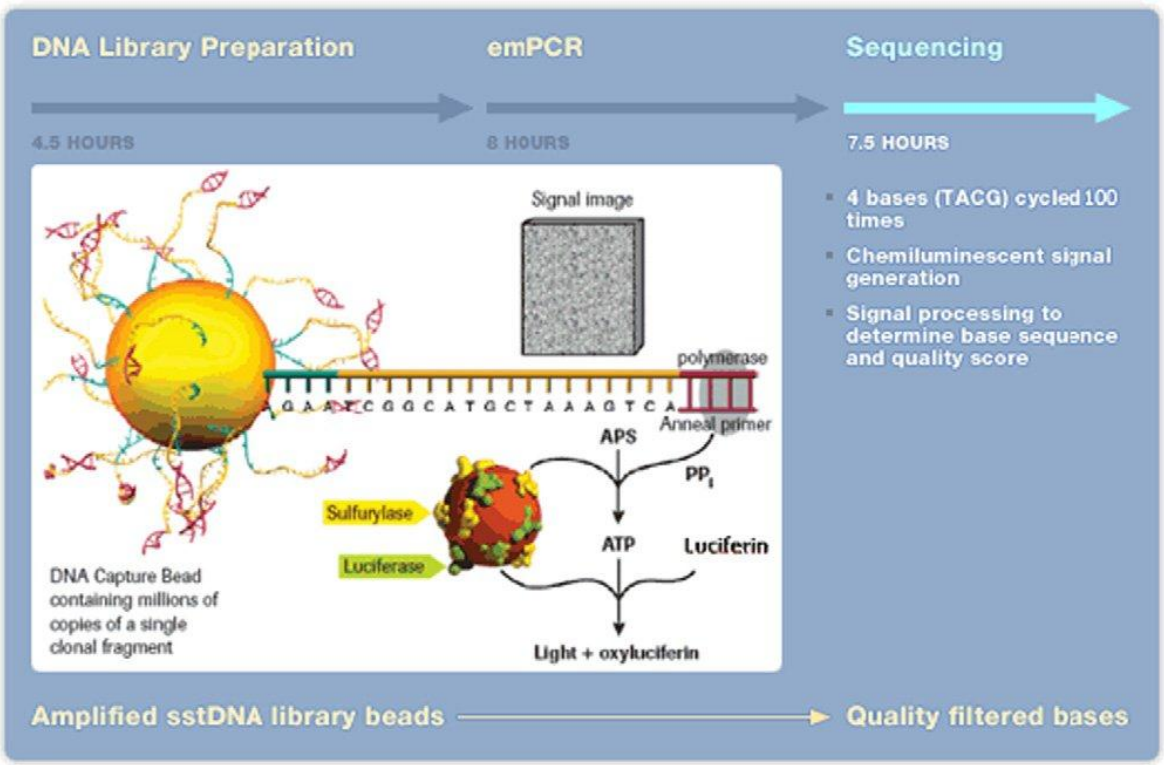

Figura 5. Pirosecuenciación (tomado de Droege y Hill, 2008).

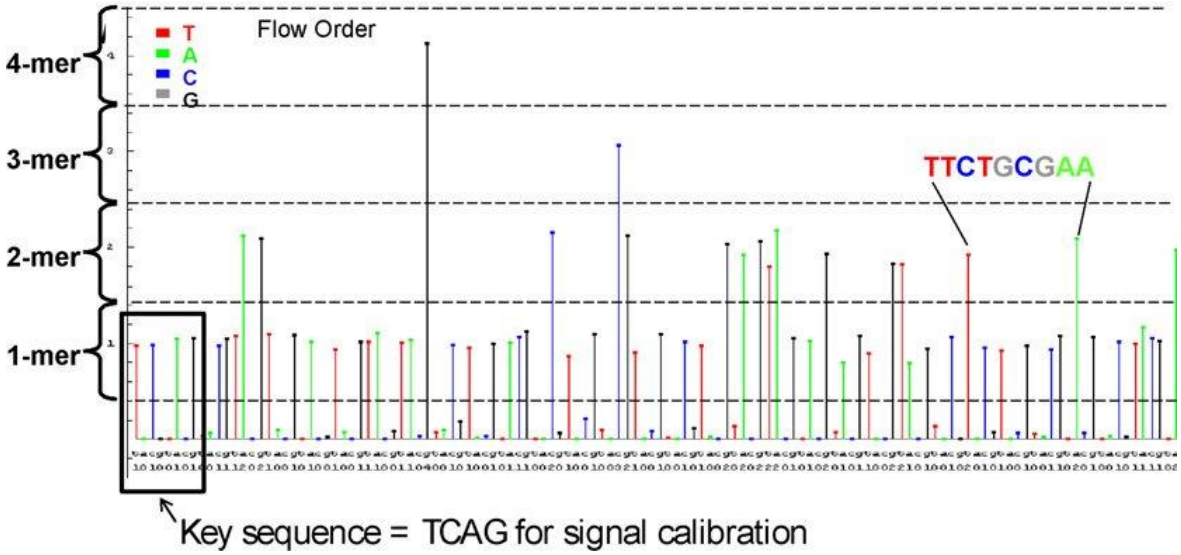

Figura 6. El flujograma. La intensidad de la señal proporcionada en el eje–y es proporcional al número de nucleótidos incorporados en un flujo de un solo nucleótido (tomado de Droege y Hill, 2008).

El secuenciador 454 GSFLX Titanium secuencia entre 400-600 millones de bases en 10 horas de corrida, permitiendo secuenciar grandes cantidades de DNA a bajo costo comparado con el método de Sanger.

La tecnología de secuenciamiento 454 también es capaz de detectar mutaciones en un grupo de amplicones con gran sensibilidad, los cuales puedan tener aplicaciones a nivel clínico especialmente en el diagnostico del cáncer y HIV. Una limitación del secuenciamiento 454 es la presencia de homopolímeros de segmentos de DNA, por ejemplo las regiones que contienen copias de una sola base (A, G, C o T). Ya que el pirosecuenciamiento se basa en la magnitud de la emisión de luz para determinar el número de bases repetidas, la llamada de las bases erróneas puede ser el problema con los homopolímeros. Otra desventaja del secuenciamiento 454 es que mientras que esta tecnología es barata y rápida considerando por lectura de cada base, cada corrida es cara y además no se recomienda para secuenciar pequeños fragmentos tales como se usan para los análisis filogenéticos.

2.10 Ensamblaje de secuencias ESTs 454 y Sanger

Con el advenimiento de nuevas tecnologías de secuenciamiento como illumina®, 454 ™, SOLID™ se genera gran cantidad de información de secuencias. Por lo tanto, se necesitan a su vez herramientas bioinformáticas que estén acorde con las nuevas tecnologías y que

posean capacidad para procesar grandes cantidades de información. En la actualidad existen programas libres y comerciales que son utilizados para el ensamblaje de secuencias. Dentro de los programas que han sido utilizados para el ensamblaje de secuencias obtenidas mediante el método de Sanger tenemos TGICL (Pertea *et al*., 2003), CAP3 (Huang y Madan, 1999) y las que son usadas para los generados por la tecnología de secuenciamiento illumina®, 454 ™, SOLID™ tenemos a MIRA (Chevreux *et al*., 2004), est2assembly (Papanicolaou *et al*., 2009), Velvet (Zerbino y Birney, 2008) Newbler (Roche Inc., Marguelies *et al*., 2005), PAVE (Soderlund *et al*., 2009) y SeqMan Ngen (DNASTAR Inc.). Estos han sido creados para el ensamblaje de secuencias derivadas de diferentes sistemas de secuenciamiento.

El SeqMan NGen™ es una aplicación que usa un único algoritmo para ensamblar las secuencias generadas por la tecnología illumina®, 454 ™, SOLID™, Sanger y los híbridos de éstas. El algoritmo de NGen se basa en secuencias únicas nucleotídicas, llamadas *mers*, que se encuentran en las regiones sobrelapadas de los *reads*. Los *mers* que son comunes a dos o más fragmentos de los *reads* son alineados para determinar a lo largo de todos los *reads*. Los *reads* que se sobrelapan tienen muchos *mers* en común pero sólo unos pocos *mers* son necesarias para identificar el *overlapping*. Estos *mers* son llamados *mer tags* (DNASTAR 2009). El poder de usar *mer tags* recae en la capacidad de NGen de seleccionar los *mers* que son más probables que se encuentren sólo una vez en la secuencia original de DNA.

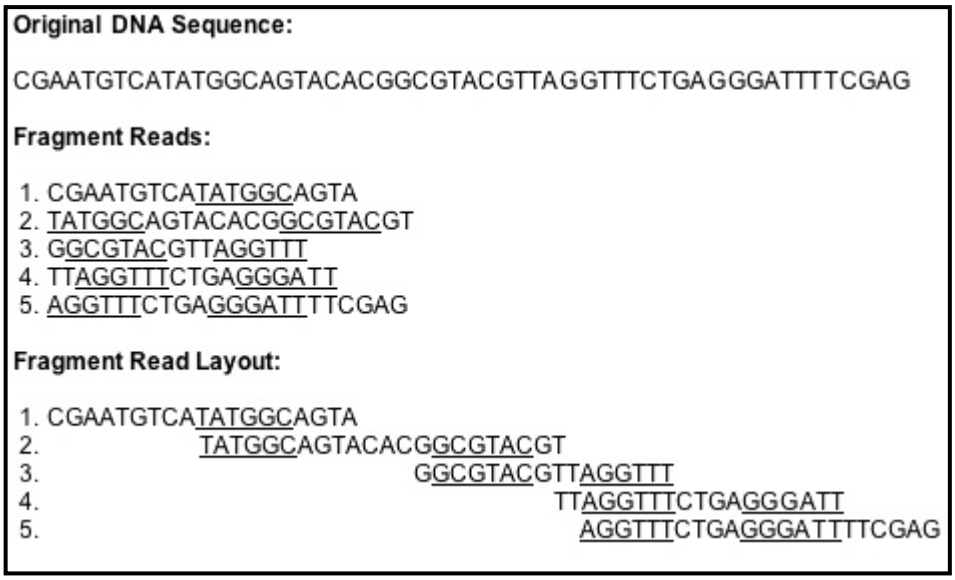

Figura 7. Usando Mer Tags para identificar *Overlaps***.**
Los *mer tags*, en este ejemplo, son el grupo de 6 nucleótidos sobrelineados.

2.11 Anotación de genes

La anotación de genes generalmente se realiza vía comparación de secuencias nucleotídicas y proteícas con genes frecuentemente derivados de otros organismos modelo, en el caso de plantas es muy común utilizar las secuencias de *A. thaliana* debido a que se tiene mayor información genómica y funcional. El consorcio del *Gene Ontology* (Ashburner *et al*., 2000), anota los genes tomando en cuenta 3 categorías: 1) la función celular o molecular, 2) el proceso metabólico en el que se encuentra involucrado y 3) la ubicación celular de esta proteína. Las comparaciones entre secuencias nucleotídicas o proteícas se han realizado generalmente utilizando la herramienta BLAST (Altschul *et al*., 1997). Una herramienta bioinformática de fácil uso y acceso en un laboratorio con bajos recursos informáticos para la atribución de funciones o localización celular es el Blast2GO (Conesa *et al*., 2005). Este programa nos permite controlar cada paso en el proceso de anotación de las secuencias. Además, este programa permite correr online el BLAST directamente contra la base de datos del NCBI o poder descargar la base de datos y correr localmente. Los pasos generales de anotación son: realizar el Blast, hacer mapeo de las secuencias con los GO términos, realizar la anotación y la validación de los mismos. En el proceso del Blast se considera los 20 primeros *hits*, los cuales serán evaluados para el proceso de anotación. Existen reglas definidas para el proceso de anotación usando el Blats2GO (Conesa *et al*., 2005)

Por otro lado, el UniRef100 es una base de datos que está disponible en la página web del Uniprot **(**ftp://ftp.uniprot.org/pub/databases/uniprot/uniref/uniref100/). UniRef100 está basado en todas las secuencias de UniProtKB, también contiene secuencias seleccionadas de Uniparc incluyendo algunas proteínas de *Bos taurus, Canis canis, Drosophila melanogaster, Takifugu rubripes, Homo sapiens, Saccaromyces cerevesiae, Mus musculus, Rattus norvegicus, Tetraodon, Xenopus y Danio rerio*, que están en la base de datos del Ensembl. UniRef100 es producido por el alineamiento y agrupamiento "*clustering*" de todos estos registros tomando en cuenta la identidad de las secuencias. Secuencias idénticas y subfragmentos de éstas son presentadas como una sola entrada en uniRef100 provenientes de varias accesiones de secuencias que han sido unidas, las secuencias proteicas del Uniref100 tienen una referencia en sus correspondientes UniProtKB.

Una base de datos proteícas es la de *Arabidopsis thaliana*, que se puede obtener de (TAIR, www.arabidopsis.org). Esta base de datos presenta un gran número de secuencias proteicas anotadas, por lo tanto sirve como una buena fuente de comparación en procesos de anotación.

2.12 Índices de genes

Las secuencias EST (*expressed sequence tag*) permiten la identificación de genes expresados. Los ESTs son secuencias parciales de cDNA y que han sido usados extensivamente en el descubrimiento de genes de muchos organismos. Las secuencias 454 derivados de cDNA también pueden ser consideradas como EST. Las colecciones de EST son altamente redundantes, ya que en las bibliotecas de cDNA generalmente existen bastantes secuencias derivadas del mismo locus. La redundancia (varias secuencias por locus más que una sola secuencia por locus) de EST es alta cuando se usa técnicas como el secuenciamiento 454 que ya da como resultado miles de *reads* derivadas del mismo locus, dependiendo de la abundancia del cDNA o del gen. El ensamblaje de EST para formar *contigs* disminuye la complejidad de una base de datos y el número de secuencias que provienen del mismo locus. La representación de los ESTs ensamblados es llamada "índice de genes". Los índices de genes consisten de ESTs ensamblados (*contigs*), más las secuencias que se mantuvieron sin ensamblar (*singletons*). Se han establecido índices de genes para un gran número de animales, plantas y otros organismos. La calidad del índice de genes depende principalmente del número de EST disponible para su ensamblaje. Para el camote existe un ensamblaje tentativo con aproximadamente 22,000 ESTs disponibles en http://plantta.jcvi.org/index.shtml (Childs *et al.*, 2007).

2.13 Microsatélites

Los microsatélites son marcadores moleculares también llamados *simple sequence repeats* (SSRs) que consisten en repeticiones pequeñas en tándem de 1-6 pb. Los SSRs pueden ser repeticiones di, tri, tetra o penta nucleotídicas. Los SSRs son marcadores útiles debido a que ellos son abundantes, hipervariables debido a su alta tasa de mutación (10^{-3}), tienen amplia distribución en los genomas y se encuentran distribuidos tanto en las regiones codificantes como no codificantes (Powell *et al*., 1996). Los microsatélites son marcadores

basados en PCR, los iniciadores están localizados en las regiones flanqueantes de los loci repetitivos, los cuales son usados para amplificar fragmentos de DNA que contengan el motivo SSR; el polimorfismo es expresado en la longitud de los fragmentos obtenidos los cuales indican las diferencias alélicas. Los marcadores microsatélites muestran una herencia mendeliana y codominancia. Los marcadores microsatélites han sido usados en estudios forenses, identificación de cultivos, la evaluación de cultivares de parentesco, análisis de la diversidad genética, los estudios evolutivos, la construcción de mapas moleculares, y para obtener las patentes y derechos de propiedad sobre las variedades de plantas (Powell *et al*., 1996, Röder *et al*., 1998, Gupta *et al*., 1999, Buteler *et al*., 2002).

Los marcadores SSRs pueden ser desarrollados por bibliotecas genómicas enriquecidas de SSR (Gianfranchesci *et al*., 1998) o si la información de las secuencias está disponible, los motivos SSR se pueden identificar utilizando un *script* apropiado (Da Maia *et al*., 2008).

El análisis genético utilizando microsatélites involucra amplificación por PCR de DNA usando iniciadores complementarios a las regiones flanqueantes de un locus, tamaño de fraccionamiento y la visualización de los productos amplificados en un gel de poliacrilamida o en un analizador de fragmentos, como es un sistema de análisis de DNA LI-COR 4300. Cada alelo presente en la muestra es representada como una banda y puede ser fácilmente registrado para el análisis estadístico. Sin embargo, en la práctica un alelo a menudo está representado por más de una banda, haciendo el registro difícil, las bandas múltiples por cada alelo se debe a "*stuttering*" presentado diferentes tamaños de fragmentos amplificados en el PCR como consecuencia del deslizamiento de la DNA polimerasa debido a repeticiones por tándem. Este efecto es más observado en motivos di, tri o tetranucleótidos de SSR y pueden mejorarse incrementando la temperatura de hibridación, cantidad de $MgCl_2$ en la reacción, disminuyendo el número de ciclos de PCR, o por la prolongación del tiempo de extensión final en el PCR o usando iniciadores marcados con fluorescencia que resulta en la detección de sólo una hebra.

3 MATERIALES Y MÉTODOS

3.1 Procedimiento en la generación del índice de genes

3.1.1 Material biológico

Plantas de camote *Ipomoea batatas* cultivar Tanzania con código CIP N° 440166.

3.1.2 Experimento de estrés hídrico

Se hizo crecer *Ipomoea batatas* (L.) Lam. cultivar Tanzania en 5 macetas por un mes en condiciones de invernadero y después se les expuso a estrés hídrico (sin irrigar) durante dos meses. Después de ese tiempo se colectó hojas y tallos por separado y se las almacenó a -70 °C hasta su uso.

3.1.3 Extracción de RNA

Se realizó la extracción de RNA a partir de 200 mg de hojas y tallos por separado de *Ipomoea batatas* (L.) Lam. cultivar Tanzania según el protocolo de TRIzol® *reagent* (Invitrogen, ver anexo1). Se cuantificó las muestras de RNA utilizando el equipo *nanodrop*. Asimismo, para corroborar la calidad del RNA, se realizó una corrida en gel de agarosa al 1% en tampón TBE a 80 voltios por 50 minutos usando como referencia el RNA ladder Gibco®.

Las muestras fueron enviadas según indicaciones de EVROGEN (http://www.evrogen.com/support/sample-requirement.shtml) el cual consistía en agregar a las muestras de RNA, acetato de sodio 3M (0.1 volumen) y 3 volúmenes de etanol al 96%.

3.1.4 Síntesis y normalización del cDNA

La síntesis y la normalización de cDNA fueron realizadas por la compañía EVROGEN. Para la síntesis de cDNA se utilizó el RNA total y se usó la metodología *SMART™ Technology* desarrollado por Zhu *et al*. (2001), que permite obtener cDNA de dos hebras en tamaño completo a partir de pequeña cantidad de RNA.

Reporte de EVROGEN de síntesis de cDNA

Para la síntesis de cDNA se utilizaron los siguientes oligonucleótidos**:** *SMART Oligo II oligonucleotide* 5'-AAGCAGTGGTATCAACGCAGAGTACGCrGrGrG-3', *CDS primer* 5'-AAGCAGTGGTATCAACGCAGAGTA-d(T)30-3', *SMART PCR primer* 5'-AAGCAGTGGTATCAACGCAGAGT-3'.

a. Síntesis del cDNA de hebra simple

La mezcla de reacción con el iniciador de "*annealing"* conteniendo (5µL): 0.3 µg de RNA total, 10 pmol SMART OligoII oligonucleotide, 10 pmol CDS primer. La mezcla de reacción fue calentada a 72°C por 2 minutos y enfriada en hielo por 2 minutos. La síntesis de la hebra simple de cDNA fue inicializada por una mezcla de RNA-iniciador con la transcriptasa reversa en un volumen final de 10 µL, conteniendo: Tampón First-strand 1X (50mM Tris-HCL (pH 8.3) 75mM KCl 6mM $MgCl_2$, 2mM DTT, 1mM de cada dNTP. La reacción de síntesis fue incubada a 42ºC por 2 horas en un incubador y luego enfriada en hielo.

b. Preparación de la doble hebra cDNA

El cDNA de hebra simple fue diluido 5 veces con tampón TE, calentado a 70ºC por 7 minutos y usado para la amplificación por PCR *Long-Distance PCR* (Barnes, 1994). La reacción de PCR (50 µL) contenía: 1 µL de cDNA diluido, tampón de reacción 1X (BD Biosciences Clontech), 200 µM dNTPs, 0.3 µM SMART PCR primer, Mezcla de 1x Advantage polymerize (BD Biosciences clontech). Se llevaron a cabo 21 ciclos de PCR en el termociclador MJ Research PTC-200. Cada ciclo incluyó 95ºC por 7 segundos, 65ºC por 20 segundos, 72 ºC por 3 minutos. Los productos de amplificación de PCR del cDNA fueron purificados usando un *kit* QIAquick PCR *purification kit* (QIAGEN, CA) y concentrado con precipitación por etanol. El sedimento de cDNA fue diluido con agua milliQ para una concentración final de 50 ng/ µL.

3.1.5 La normalización del cDNA

La normalización del cDNA se realizó de acuerdo al método de Zhulidov *et al*. (2004). Éste método involucra la desnaturalización y reasociación, la degradación de cDNA de doble hebra y la amplificación por PCR de los cDNA normalizados (cDNA de hebra

simple que se encuentran en número de copias similares). El elemento crucial en este método es el uso de la enzima *Duplex-specific nuclease* (DSN) descubierta por Shagin *et al*. (2002), esta enzima presenta una afinidad por los cDNAs de doble hebra o híbridos DNA-RNA, que se forman debido a los transcriptos que se encuentran en abundancia en el transcriptoma. De esta manera, al degradar las secuencias abundantes se reduce el número de copias de éstas en la biblioteca y quedan presentes en un número de copias equivalente a los demás transcriptos presentes en la biblioteca de cDNA.

Reporte por EVROGEN de la normalización de cDNA

a. Hibridización

La reacción de hibridización contenía: 3 µL (cerca de 150 ng) de cDNA de doble hebra purificada, 1 µL de tampón de hibridización 4X (200 mM HEPES-HCL, pH 8.0; 2M NaCl). La mezcla de reacción fue superpuesta con una gota de aceite mineral e incubado a 98 ºC por 3 minutos y 68 ºC por 5 horas.

b. Tratamiento con la enzima Duplex-specific nuclease (DSN)

Los siguientes reactivos precalentados fueron agregados a la reacción de hibridización a 68 ºC: 3.5 µL agua milliQ, 1 µL de tampón DNAsa 5X (500 mM Tris-HCL, pH 8.0; 50 mM $MgCl_2$, 10 mM DTT), 0.5 µL de enzima DSN. Luego, la incubación fue prolongada a 67°C por 20 minutos. Tras el término del tratamiento con DSN, la enzima DSN fue inactivada mediante calentamiento a 97 °C por 5 minutos.

c. Amplificación del cDNA normalizado

Previamente la muestra de cDNA fue diluida por la adición de 30 µL de agua milliQ y ésta fue usada en la amplificación por PCR. La reacción de PCR (50 µL): 1 µL de cDNA diluido, tampón de reacción 1X *Advantage* 2 (BD Biosciences Clontech®), 200 µM dNTPs, 0.3 µM SMART PCR primer, mezcla de polimerizar 1x Advantage 2(BD Biosciences Clontech). El PCR fue llevado a cabo en el termociclador MJ Research PTC-200. Se realizaron 18 ciclos de PCR y cada ciclo de PCR incluyó: 95ºC por 7 segundos, 65ºC por 20 segundos, 72 ºC por 3 minutos.

3.1.6 Secuenciamiento con la tecnología 454

El secuenciamiento de las bibliotecas de cDNA normalizadas de hojas y tallos fue realizado en la Universidad de Liverpool, Reino Unido. Se realizó un "*quarter run*" de una corrida de 454 FLX para la biblioteca de hojas y otro "*quarter run*" de una corrida 454 FLX TITANIUM para la biblioteca de tallo.

3.1.7 Análisis *in silico*

Una vez recibida las secuencias se realizó el trabajo *in silico*, para lo que fue necesario el uso de sistemas computacionales de uso más sofisticado, es así que se trabajó con dos tipos de servidores y programas bioinformáticos especializados. En el trabajo también se utilizó bases de datos de secuencias nucleotídicas y proteícas e índice de genes de otras plantas.

Servidor de Linux

La computadora que se usó presentó las siguientes características: cuatro nodos y cada nodo cuenta con dos procesadores AMD Opteron(tm) Processor 248 - 64 Bits, cpu MHz: 2205.072, Memoria RAM: 4GB, Sistema Operativo: Rocks Linux 4.2 – 64 Bits.

Servidor de Windows

El servidor presentó las siguientes características, cuatro procesadores INTEL XEON CPU 3.00 GHz – 64 bits, Memoria RAM 8 GB, Sistema Operativo: Windows XP 64 bits.

Programas informáticos

- SSR locator: http://www.ufpel.edu.br/~lmaia.faem.
- Bioedit 7.0.9.1: http://www.mbio.ncsu.edu/BioEdit/bioedit.html
- Notepad++ portable: http://notepad-portable.softonic.com/
- SeqMan Ngen (DNAstar): DNASTAR, Inc., Madison USA
- Lasergen (DNAstar): DNASTAR, Inc., Madison USA
- Blast2GO: www.Blast2go.org

- Scripts (Blast2tableformat8, Chomper, partidor, selectedv02) CIP, RIU, Lima, Perú.
- Putty.exe: http://www.chiark.greenend.org.uk/~sgtatham/putty/download.html
- winscp417.exe:http://www.filewatcher.com/m/winscp417.exe.1305600.0.0.html
- Blast del NCBI version 2.2.14.:http://Blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastNews#1
- SeqClean: http://compbio.dfci.harvard.edu/tgi/software/
- Repeat masker: http://www.repeatmasker.org/
- CAP3: http://seq.cs.iastate.edu/

Bases de datos

- Uniprot uniRef100: http://www.ebi.ac.uk/uniref/
- TIGR Plant Gene Assembly: http://plantta.jcvi.org/index.shtml
- TIGR Ipomoea nil Gene Index: http://compbio.dfci.harvard.edu/cgi-bin/tgi/gimain.pl?gudb=morning_glory
- Plant Gene Database: http://www.plantgdb.org/
- NCBI Genbank: http://www.ncbi.nlm.nih.gov/genbank/

3.1.7.1 Procesamiento y limpieza de las secuencias

La limpieza de las secuencias 454 de las bibliotecas de hojas y tallos y de los ESTs del Genbank se realizaron en el servidor en Linux High Performance Computer (HPC) perteneciente al Centro Internacional de la Papa (http://hpc.cip.cgiar.org/). Se bajaron las secuencias EST del NCBI de (http://www.ncbi.nlm.nih.gov/sites/entrez). Se utilizó el SeqClean para la eliminación de las secuencias de vectores de las secuencias EST del Genbank. Para las secuencias 454 también se utilizó el SeqClean para eliminar los adaptadores que se usaron en la creación de las bibliotecas (SMART primers). Las secuencias repetitivas y las colas poly-A fueron enmascaradas con el programa RepeatMasker 3.2.7 (http://www.repeatmasker.org/RMDownload.html). Aquellas secuencias con menos de 100 bp fueron eliminadas antes del ensamblaje.

3.1.7.2 Ensamblaje de las secuencias

Las secuencias obtenidas mediante la tecnología 454 y EST del NCBI del Genbank fueron ensambladas juntas utilizando el programa SeqMan NGen, (DNASTAR, Madison, WI, USA). Se realizó un ensamblaje *de novo*. Además, se realizó varios ensamblajes con la modificación del parámetro "*MinMatchPercent*" desde 70, 75, 80, 85 y 90%, con la finalidad de obtener una optimización del ensamblaje. Esto se realizó para determinar el mejor ensamblaje, considerando un mínimo de redundancia y un máximo de representación de diferentes miembros de familias de genes como indicador de un buen ensamblaje. Los parámetros que se tomaron en cuenta para el ensamblaje se presentan en la Tabla 1

Tabla 1. Parámetros considerados para el ensamblaje de las secuencias ESTs de *Ipomoea batatas* (L.) Lam

setParam CoverageType: genoma setParam GapPenalty: 7 (penalidad por cada espacio introducido) setParam GenomeLength: 4500000 (longitud hipotética del genoma/transcriptoma) setParam Max454SeqLen: 600 (longitud máxima de un *read*) setParam MatchSize: 21 (longitud de un *mer tag*) setParam MatchSpacing: 40 (espacio entre los *mer tags*) setParam Min454SeqLen: 40 (longitud mínima para un *read*) setParam MinMatchPercent: 80 (porcentaje mínimo de identidad en los sobrelapamientos) setParam MismatchPenalty: 12 (penalidad por "mismatch")

3.1.7.3 Minería de datos

Para la mayoría del proceso y análisis de las secuencias se ha usado la herramienta Blast, con los parámetros por *default* (Altschul *et al*., 1997).

BLASTn

Se realizó Blastn considerando un E-value de 10^{-6} a las secuencias del índice de genes contra las secuencias EST de *Ipomoea batatas* e *Ipomoea nil* ensambladas por TIGR (http://plantta.jcvi.org/cgi-bin/plantta_release.pl).

BLASTx

Se realizó BLASTx tomando como E-value de 10^{-6} de las secuencias del índice de genes contra las secuencias de proteínas de *Arabidopsis thaliana* (TAIR, www.arabidopsis.org). Éste se usó para seleccionar el mejor ensamblaje. Además, se realizó otro Blastx de las secuencias del índice de genes contra las secuencias UniRef100 de la base de datos del uniprot (http://www.uniprot.org/, Suzek *et al*., 2007).

3.1.7.4 Anotación de secuencias

La anotación de genes al nivel de ontología de genes se realizó con el programa Blast2GO (Conesa *et al*., 2005, Ashburner *et al*., 2000). El Blast2GO es una herramienta bioinformática que permite realizar el Blastx con la base de datos del NCBI o el Swissprot, mapear las secuencias obtenidas del Blastx con los términos GO de la base de datos del gene ontology, anotar las secuencias. Adicionalmente, nos permite realizar anotación con la base de datos del Interpro y el KEGG.

Para el proceso de anotación de los genes de camote se siguió los siguientes pasos: 1) se realizó el Blastx contra la base de datos nr del NCBI con un E-value de 10^{-3}, pero para la anotación de las secuencias se consideró sólo los *hits* con un E-value de 10^{-6}; 2) se realizó la validación de la anotación y 3) adicionalmente se realizó la búsqueda de las anotaciones de las enzimas encontradas en la base de datos *Kyoto Encyclopedia of Genes and Genomes* (KEGG).

3.1.7.4.1 Identificación de los ORF (open reading frames)

Se realizó la identificación de los ORF en los *contigs* y *singletons* que no tuvieron *hits* con la base de datos de proteínas UniRef100, Para ello se uso el programa Orf-Predictor (http://proteomics.ysu.edu/tools/OrfPredictor.html). Esto se realizó con la finalidad de poder determinar si las secuencias sin *hits* representan posibles genes y no artefactos como producto del proceso de ensamblaje.

3.2 Procedimiento para el desarrollo de los microsatélites

3.2.1 Identificación y diseño de iniciadores microsatélites

Con la finalidad de identificar nuevos marcadores microsatélites, para el diseño de iniciadores se excluyeron las secuencias similares a las secuencias EST del Genbank, que fueron usados para los microsatélites previamente publicados (Hu *et al*., 2004). Para ello se realizó un Blastn de nuestras secuencias contra las secuencias del Genbank considerando un E-value de 10^{-30} y además sólo se consideró los *contigs* y *singleton* únicos o grupo de secuencias homólogas que fueron probablemente derivadas del mismo locus, de esta manera se evitó la redundancia de los iniciadores. Estos iniciadores fueron posteriormente usados en la evaluación de los 8 genotipos de camote.

La identificación y el diseño de los microsatélites se realizó utilizando el Perl script SSR locator (Da Maia *et al*., 2008). Se consideraron los siguientes parámetros para la búsqueda de los motivos repetitivos, como mínimo 7 dímeros, 5 trímeros y 5 tetrámeros. Para el tamaño del amplicón se consideró entre 100 a 200 pares de bases, para el tamaño del iniciador se tomó como tamaño mínimo y máximo 20 y 25 pares de bases respectivamente y la temperatura de *melting* de los iniciadores fue de 55 ºC como mínimo y 65 ºC como máximo y para el contenido de GC se tomo como mínimo 40 y óptimo 50.

3.2.2 Amplificación de los microsatélites

Las amplificaciones de los microsatélites fueron evaluados en 8 genotipos, que consistieron de seis accesiones hexaploides de *I. batatas* y dos diploides de *I. trifida*. Cuatro de éstas representan clones de Asia, Latinoamérica y África, y las restantes cuatro son progenitores provenientes de Uganda (Tanzania), USA (Beauregard) y Panamá (*Ipomoea trífida*) (Tabla 2). La extracción de DNA fue realizada de acuerdo al método de Doyle y Doyle (1987). La mezcla de reacción fue realizada en un volumen de 10 µL y contenía los siguientes componentes: tampón PCR 1X con 2.5 mM $MgCl_2$ (New England Biolabs®), 0.2 mM de cada dNTPs (Invitrogen®), 22 pM primer *forward* (IDT®), 15pM primer *reverse* (IDT®) y 25 pM de M13Forward 700/800 IRDye (IDT®) y 0.25 U *Taq* polimerasa (New England Biolabs®) y se tomó aproximadamente 30 ng de DNA genómico. El PCR fue realizado en un termociclador PTC-100 o PTC-200 (MJ Research Inc.). Las condiciones de amplificación fueron 4 minutos para la desnaturalización,

seguido de 30 ciclos de PCR de 1 minuto a 94 °C, 1 minuto para la temperatura de *annealing* (ver en la tabla 6), 1 minuto a 72 °C, seguido de una síntesis de terminación a 72 °C por 7 minutos. Se le adicionó 5uL de la solución azul stop de PCR antes de cargar la muestra. Los fragmentos obtenidos de la amplificación fueron separados en geles de poliacrilamida al 6% por el sistema Li-Cor 4300 (LI-COR Biosciences NE, USA). Se utilizó un marcador de 50-350pb (LI-COR, USA) para determinar el tamaño molecular y los geles fueron analizados tomando en cuenta el método de Ghislain *et al*. (2009).

Tabla 2. Lista de accesiones de camote usado para la evaluación de los microsatélites

Código CIP	Especie	Nombre del cultivar	Origen	Descripción
107665.9	*Ipomoea trifida*	M9	CIP	diploide, padres de mapeo
107665.19	*Ipomoea trifida*	M19	CIP	diploide, padres de mapeo
401206	*Ipomoea batatas*	401206	Mexico	hexaploide, variedad local
420027	*Ipomoea batatas*	Zapallo	Peru	hexaploide, variedad local
440025	*Ipomoea batatas*	Xushu 18	China	hexaploide, variedad mejorada
440131	*Ipomoea batatas*	Naveto	Papua New Guinea	hexaploide, variedad local
440132	*Ipomoea batatas*	Beauregard	USA	hexaploide, padres de mapeo, variedad mejorada
440166	*Ipomoea batatas*	Tanzania	Uganda	hexaploide, padres de mapeo, variedad local

4 RESULTADOS

4.1 Obtención del índice de genes

4.1.1 Obtención de RNA

Se obtuvo las muestras de RNA total de hojas y tallos de *Ipomoea batatas* (L.) Lam. a partir de 200 mg (Figura 8, Figura 9). El tamaño de los fragmentos de RNA estuvo en un rango de 0,3 y 3 kb que fue determinado utilizando el marcador de RNA Gibco® (Figura 8). La concentración de la solución de RNA obtenida de las hojas y tallos fue de aproximadamente 1.5µg/µL.

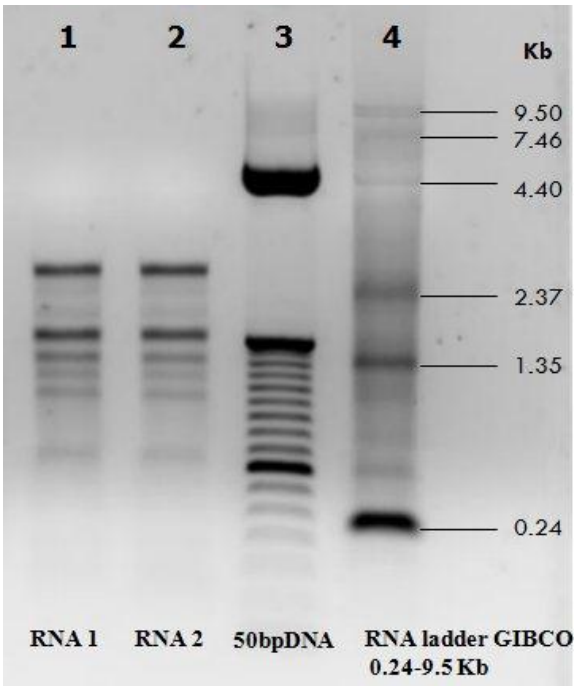

Figura 8. Extracción de RNA de hojas de camote.
Carril 1 y 2: Extracción de RNA camote, carril 3: marcador 50pb DNA, carril 4: marcador RNA Gibco ® 0.24-9.5.

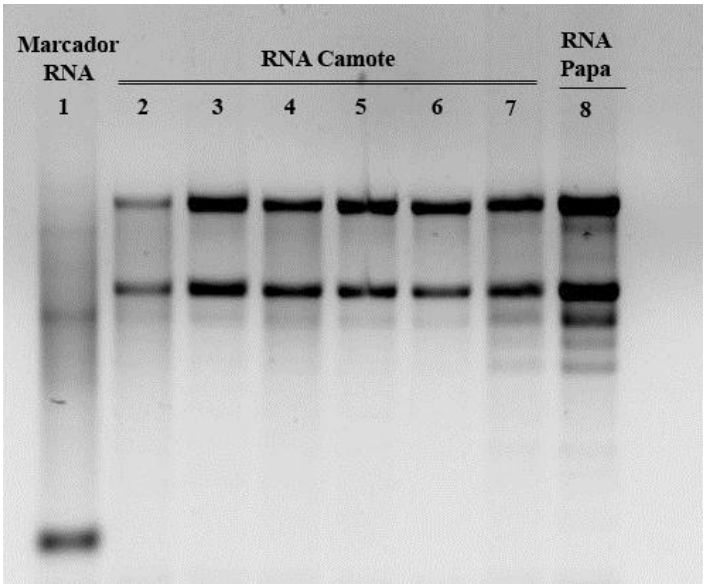

Figura 9. Extracción de RNA de tallo de camote.
Carril 1: marcador RNA gibco ® 0.24-9.5, carril 2-7: extracción RNA de camote y carril 8: extracción RNA de papa.

4.1.2 La síntesis del cDNA y normalización de las bibliotecas de cDNA

La síntesis y la normalización de la biblioteca de cDNA tanto de hojas como de tallos se realizaron por separado. En la Figura 10 se muestra el cDNA de hojas antes y después del proceso de normalización de cDNA, en el carril 1 se ve claramente una disminución en la intensidad de las bandas comparado con el carril 2, verificándose así un existoso proceso de normalización.

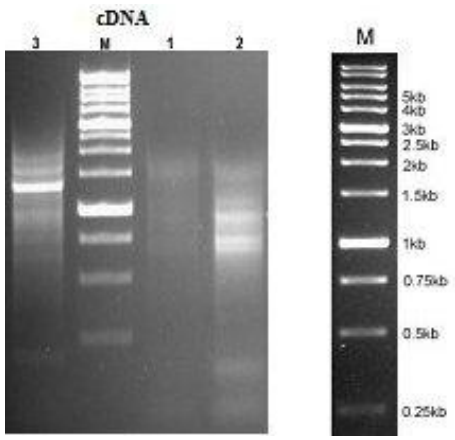

Figura 10. Normalización del cDNA de las hojas de camote.
Carril 1: cDNA normalizado, carril 2: cDNA no normalizado, carril 3: cDNA control (no normalizado).

4.1.3 Secuenciamiento 454

Se ha obtenido 87,307 *reads* que comprenden 21’292,096 bases para la corrida de 454 FLX de la biblioteca de cDNA de hojas normalizadas de plantas expuestas a sequía de camote del cultivar Tanzania, y para la biblioteca de cDNA de tallos se obtuvieron 436,817 *reads* que comprenden 136’844,411 bases. El tamaño promedio de longitud de los *reads* fue de 243.9 bases para la corrida 454 FLX y 313,3 bases para la corrida 454 FLX TITANIUM.

Los *reads* contenían adaptadores que fueron usados tanto para la síntesis de las bibliotecas de cDNA como su secuenciamiento mediante la tecnología 454. Estos adaptadores fueron eliminados al igual que los *reads* que presentaron un tamaño menor o igual a 100 bases y las secuencias de baja calidad. Además, se eliminaron las secuencias de DNA mitocondrial y cloroplástico que también se encontraron en la biblioteca. En total, se eliminaron 121,618 *reads*, quedando así 402,506 secuencias limpias que fueron usadas para el ensamblaje.

Para tener una idea de cuantos *reads* son secuencias nuevas en el índice de genes de camote y que no estuverian contenidos en las secuencias EST del Genbank, se procedió a realizar un Blastn con un E-value de 10^{-10} de las 402,506 secuencias limpias contra las secuencias EST del Genbank que previamente fueron limpiadas. Se encontró que el 31.6% del total de los *reads* tienen un *hit* considerando una cobertura e identidad de las secuencias mayor al 80%, sugiriendo que aproximadamente una tercera parte de los *reads* corresponde a genes ya conocidos para camote y el 68.4 %, que representa las dos terceras partes de los *reads*, son secuencias nuevas.

4.1.4 Ensamblaje de secuencias

Los 402,506 *reads* fueron ensamblados juntamente con los 22,094 ESTs del Genbank. Para optimizar el proceso de ensamblaje se realizó una variación en el parámetro de ensamblaje, se varió el MMP (*minimun match percent*) de 70 a 90%. En los diferentes ensamblajes se evaluó representatividad de los genes y la redundancia del índice de genes.

Se obtuvieron un diferente número de *contigs* y *singletons* como producto de la variación del parámetro MMP. Tanto el número de *contigs* como de *singletons* tuvieron un incremento linear del 70 al 80 % MMP y a partir del 80% el número de *contigs* tuvo un incremento notorio y el número de *singletons* aumentó aún mucho más (Figura 11).

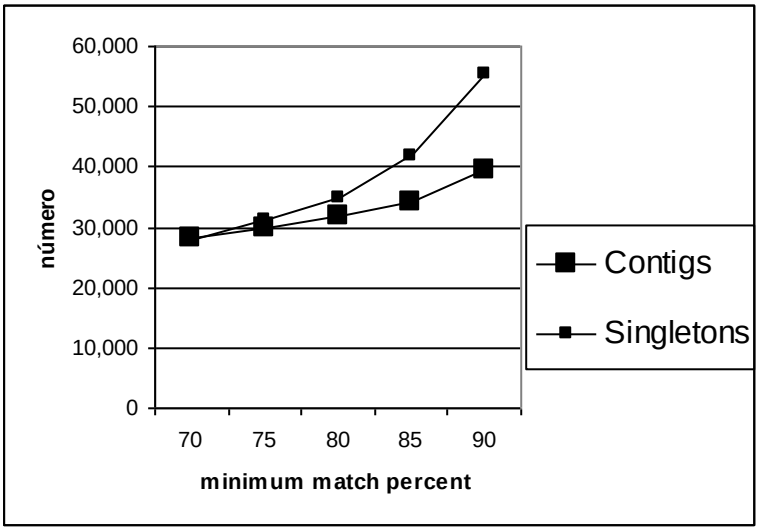

Figura 11. Variación del parámetro de ensamblaje (MimMatchPercent) de 70 a 90%. Número de *Contigs* y *Singletons* obtenidos.

La redundancia del índice de genes fue evaluada realizando un análisis de las secuencias con el Megablast y Blastx con las bases de datos proteícas de *A. thaliana*. A nivel nucleotídico se realizó *self-megablast*, por el cual se evaluó la similaridad entre los *contigs* y *singletons* del índice de genes. Las secuencias que compartían un 80% de identidad sobre el 80% de su longitud con otras secuencias fueron consideradas como redundantes. Se observó que el número de self-megablast aumentó con el incremento de MMP, indicando que las variantes alélicas se mantuvieron sin ensamblar o fueron resueltos en diferentes *contigs*. Además, se puede observar que el número de *self-megablast* se incrementa notoriamente sobre el 80% MMP (Figura 12).

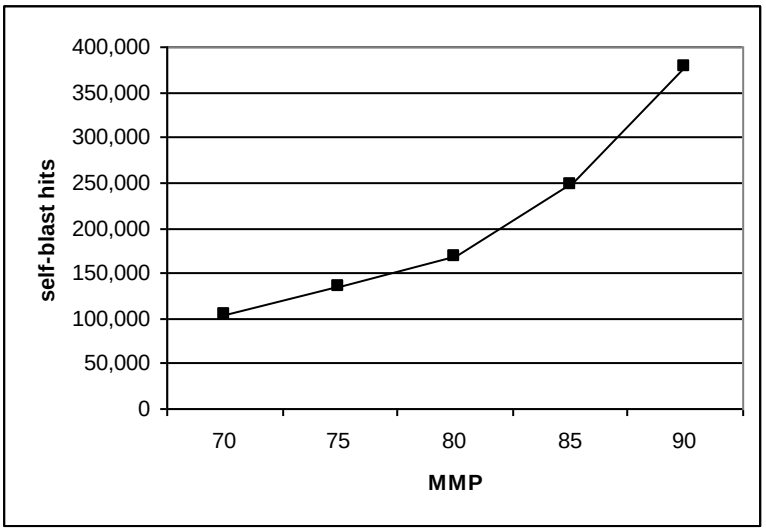

Figura 12. Variación del parámetro de ensamblaje (MimMatchPercent) de 70 a 90%. Número de *"self-blast hits".*

Adicionalmente, se evaluó la redundancia de los ensamblajes producidos a diferentes MMP realizando Blastx contra las secuencias proteícas de *A. thaliana*. Se contó el número total de secuencias y secuencias únicas obtenidas para los *contigs* y *singletons*. Se observó que el número total de secuencias se incrementó con el aumento del MMP y este incremento se hizo más notorio por encima del 80% MMP, mientras que las secuencias únicas se incrementaron linealmente en todo el rango del MMP, pero en menor medida que el número total de secuencias. Se observó que el número de secuencias únicas en los *singletons* aumentaba ligeramente a mayores valores de MMP, debido a que las secuencias permanecían sin ser ensambladas (Figura 13).

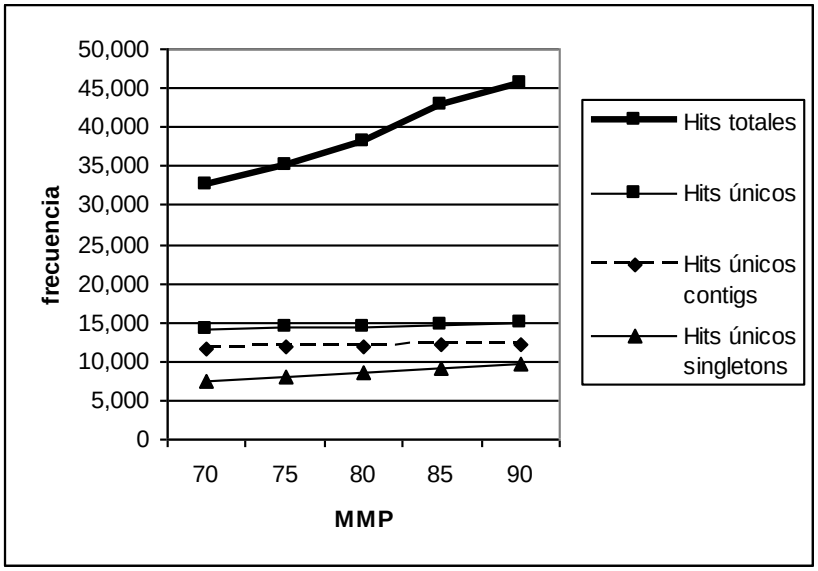

Figura 13. Variación del parámetro de ensamblaje (MimMatchPercent) de 70 a 90%. Número de *Blastx-hit* con el proteoma de *A.thaliana*.

El 70% MMP sugiere el parámetro que nos da menos redundancia, pero aparentemente une las secuencias parálogas. Entre 80 a 85% MMP la redundancia se incrementa mucho más que entre 75 y 80% MMP, por lo tanto se escogió 80% porque si bien es cierto presenta el riesgo de considerar algunas secuencias parálogas, presenta una menor redundancia que los ensamblajes a valores más altos de MMP.

4.1.5 Evaluación de la redundancia y de la representatividad en familias de genes seleccionadas

Para investigar si el ensamblaje a bajos valores de MMP une las secuencias parálogas en un solo *contig*, hemos analizado el número total y el número de miembros diferentes de 20 familias de genes escogidos al azar en los diferentes MMP (Figura 14 y Figura 15). Se realizó una búsqueda de las secuencias utilizando los identificadores y la anotación realizada de las secuencias tomando en cuenta la base proteica de *A. thaliana*. Se escogió secuencias anotadas como: la proteína dedos de Zinc A20 asociada a estrés que contiene el dominio AN1, proteína endotransglucosilasa-xyloglucano/hidrolasa, α-tubulina, β-tubulina, factor de transcripción MYB, factor de transcripción bHLH, tioredoxin, sintaxina, superóxido dismutasa, transportador de sulfato, intercambiador de sodio e hidrógeno, factor de transcripción *scarecrow*, serina carboxipeptidasa, actina, ferritina, familia de proteínas PRA1, enolasa, β-amilasa, factor de transcripcion bZIP y glutaredoxina. Los

datos muestran un incremento de 734 (70 MMP) a 1,100 (90 MMP) en el número de total de las 20 familia de genes. El número de diferentes miembros dentro de la familia de genes varía muy poco en los diferentes MMP, de 286 a 296, indicando que existe el riesgo de unir secuencias parálogas en un *contig* a bajos MMP, pero sigue siendo bajo. Por lo tanto, considerando una buena representatividad y una baja redundancia y al mismo tiempo evitando el riesgo de contener *contigs* que provengan de uniones de secuencias parálogas en vez de secuencias homólogas, se escogió un ensamblaje de mediana exigencia a 80 MMP para el índice de genes de camote. Considerando que el ensamblaje a 80 MMP todavía tiene una baja redundancia, pero una mejor representatividad de las secuencias que a menores valores de MMP, se puede ver porque existen 400 más secuencias únicas a 80MMP que a 70MMP (Figura 14 y Figura 15).

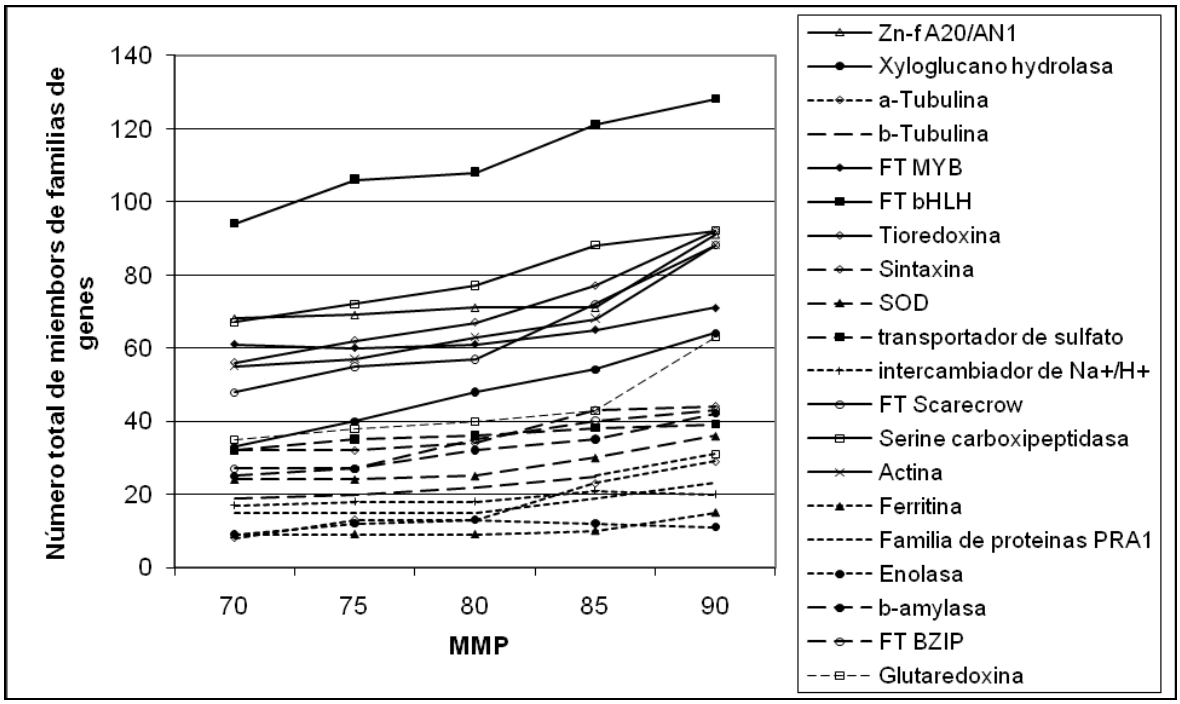

Figura 14. Número total de miembros de la familia de genes en el índice de genes de camote a diferentes MMP.
Veinte familia de genes seleccionados (dedos de Zinc A20 proteína asociada a estrés que contiene el dominio AN1, proteína endotransglucosilasa xyloglucano/hidrolasa, α-tubulina, β-tubulina, factor de transcripción MYB, factor de transcripción bHLH, tioredoxina, sintaxina, superóxido dismutasa, transportador de sulfato, intercambiador de sodio e hidrógeno, factor de transcripción *scarecrow*, serina carboxipeptidasa, actina, ferritina, familia de proteínas PRA1, enolasa, β-amilasa, factor de transcripcion bZIP, glutaredoxina).

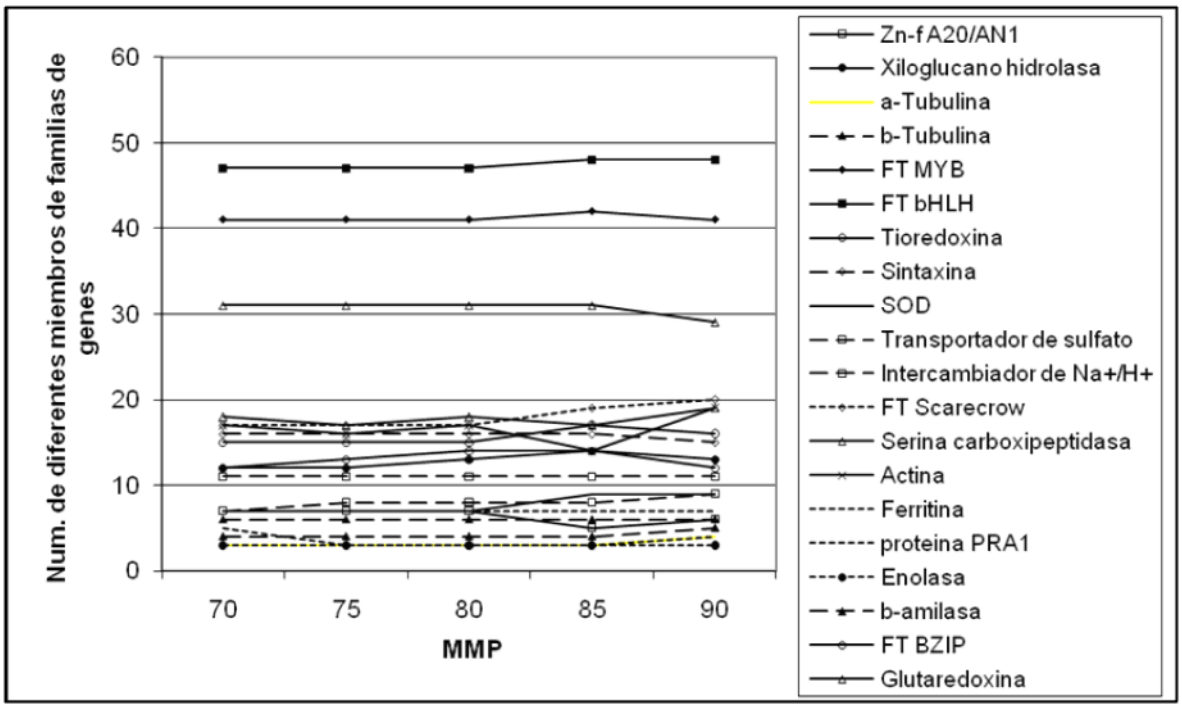

Figura 15. Número de miembros diferentes de la familia de genes en el índice de genes de camote. Veinte familias de genes seleccionados (dedos de Zinc A20 proteína asociada a estrés que contiene el dominio AN1, proteína endotransglucosilasa xiloglucano/hidrolasa, α-tubulina, β-tubulina, factor de transcripción MYB, factor de transcripción bHLH, tioredoxina, sintaxina, superoxido dismutasa, transportador de sulfato, intercambiador de sodio e hidrógeno, factor de transcripción *scarecrow*, serina carboxipeptidasa, actina, ferritina, familia de proteínas PRA1, enolasa, α-amilasa, factor de transcripcion bZIP, glutaredoxina).

A altos MMP, el número total de miembros de familia de genes se incrementa fuertemente debido a un aumento de la redundancia en el índice de genes, mientras que la representación de diferentes miembros de la familia de genes del índice de genes se mantiene casi constante a altos valores de MMP.

4.1.6 Índice de genes de camote

El ensamblaje híbrido de los *reads* 454 con las secuencias Sanger obtenidas del Genbank a 80 MMP produjo 31,685 *contigs* y 34,733 *singletons*. Los *contigs* están en un rango de 100 a 6,872 pb, con un tamaño promedio de 790 pb. La longitud acumulada de los *contigs* fue de 25’048,392 pb. La distribución del tamaño de los *contigs* se muestra en la Figura 16. Un número considerable de *contigs* largos fue obtenido; 21,929 fueron mayores a 500 pb y 8,107 a 1 kb. La cobertura del secuenciamiento tuvo un rango de 2 a 1,863 *reads* por *contigs*, con un promedio de cobertura de 12.3 (Figura 17).

El índice de genes está disponible *on-line* a:
https://research.cip.cgiar.org/confluence/display/SPGI/Home

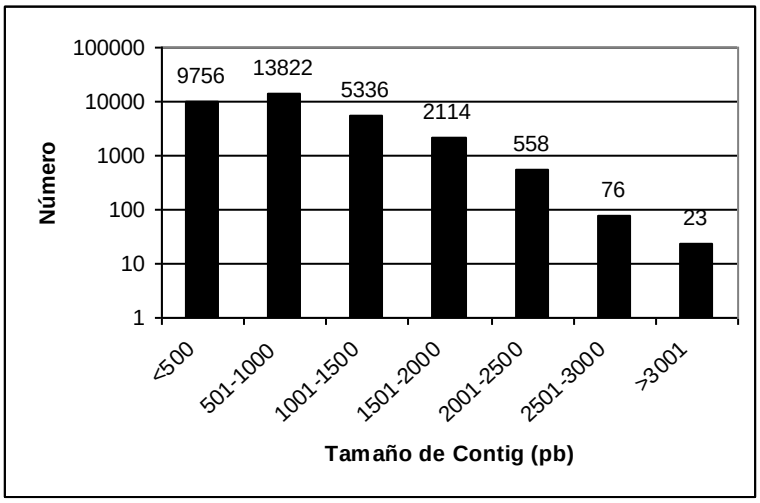

Figura 16. Distribución del tamaño de *contigs* al 80% MMP.

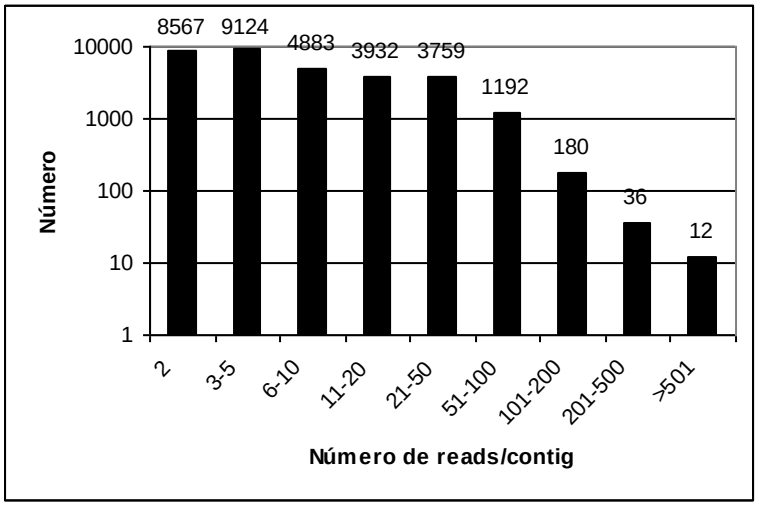

Figura 17. Número de *reads* por *contigs* al 80% MMP.

4.1.7 Comparación del índice de genes de camote con los ensamblajes de *I. batatas* e *I. nil* del TIGR

Con el propósito de realizar una comparación de nuestro ensamblaje del índice de genes de camote, se utilizó otros ensamblajes previamente realizados por el TIGR (http://plantta.jcvi.org/cgi-bin/plantta_release.pl, Childs *et al*., 2007). La comparación con este ensamblaje nos da un estimado de la cantidad de secuencias nuevas en el índice de genes de camote. Se esperó que nuestro ensamblaje incluyera todos los *contigs* y *singletons*

obtenidos en el ensamblaje usado por el TIGR, ya que estos también fueron considerados en nuestro ensamblaje. En efecto, todos menos 45 de 9,167 secuencias del ensamblaje del TIGR tuvieron *hits* con nuestro índice de genes, 22 de las 45 secuencias corresponden a secuencias de *mRNA* del Genbank. Por lo tanto, estas secuencias no fueron incluidas en el proceso de ensamblaje, ni las 23 secuencias restantes correspondientes a ESTs del Genbank de baja calidad o secuencias altamente repetitivas que fueron eliminadas en la limpieza de las secuencias antes del ensamblaje. Existen 23,151 secuencias de nuestro índice de genes que presentan similitud con 9,166 *contigs* y *singletons* del ensamblaje de TIGR. Esto indica que la redundancia en nuestro índice de genes es 2.5 veces más que el hallado en el ensamblaje del TIGR. Además, existen 43,267 secuencias de nuestro índice de genes que no tuvieron *hits* con el ensamblaje del TIGR; por lo tanto, éstas representan secuencias nuevas (Figura 18). De la misma forma, sólo 33 secuencias del ensamblaje del plantGDB de camote no estuvieron contenidas en nuestro índice de genes (dato no mostrado). Considerando que nuestro índice de genes posee una redundancia de 2.5 veces, el presente índice de genes adiciona más de 17,000 nuevas secuencias a las 9,166 secuencias previas de camote de los ensamblajes de EST. Las secuencias del índice de genes de camote también se compararon con las secuencias de *Ipomoea nil*, que es una especie cercana del camote. El índice de genes establecido de *I. nil* está conformado por 61,199 EST y 133 ETs (transcriptos maduros) resultando en 11,754 *contigs*, 9,721 EST-*singletons* y 39 ET-*singletons* (http://compbio.dfci.harvard.edu/cgi-bin/tgi/gimain.pl?gudb=morning_glory, Quackenbush *et al*., 2001). En el presente trabajo el índice de genes de camote tiene 37,986 *hits* con el ensamblaje del índice de genes de *I. nil* considerando un E-value de 10^{-6}, mientras que 28,432 secuencias no tuvieron ningún *hit*, por lo tanto estas secuencias se pueden considerar como especificas de camote. Realizando un Blastn recíproco pudimos observar que 4,146 secuencias entre *contigs* y *singletons* (19.3% de las secuencias de *I. nil*) no tuvieron *hit* con nuestras secuencias de camote, estos resultados nos indican que nuestro índice de genes tiene una buena cobertura.

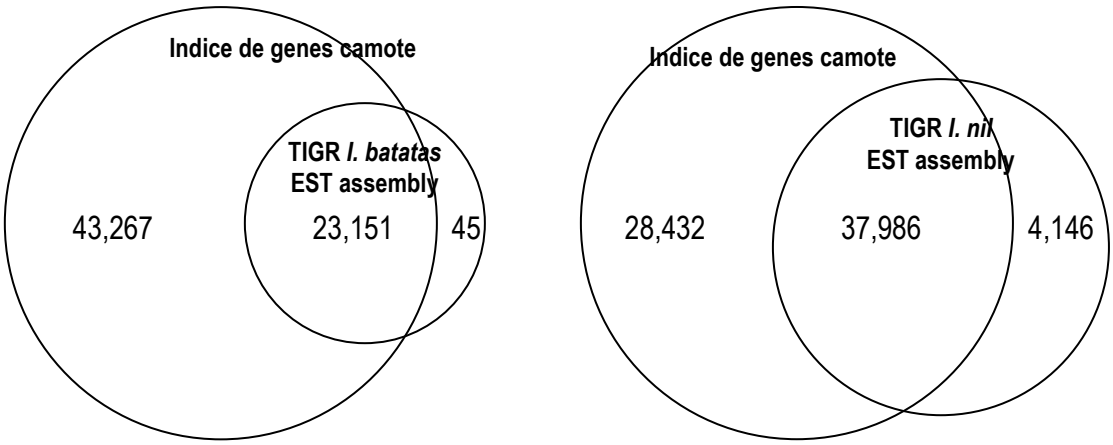

Figura 18. Comparación del índice de genes con respecto a los ensamblajes *I. batatas* e *I.nil* del TIGR. Diagrama de Venn mostrando un overlap entre el índice de genes de camote y los ensamblajes de *I.batatas* e *I.nil*.

Además, se ha evaluado el número de miembros de algunas familias de genes seleccionadas posee el índice de camote comparado con las secuencias de *A. thaliana*, cuya lista se presenta en la Tabla 3.

Tabla 3. Comparación de proteínas encontradas entre *I.batatas* y *A. thaliana*.

Familia de proteínas	***I. batatas***	***I. thaliana***
Zn-f A20/AN1	7	14
Xyloglucano/hidrolasa	13	30
α-Tubulina	3	5
β-Tubulina	4	8
Factor de transcripción MYB	41	143
Factor de transcripción bHLH	47	140
Tioredoxin	15	31
Sintaxina	16	25
Superóxido dismutasa	7	14
Intercambiador de Na+/H+	8	6
Factor de transcripción scarecrow	17	28
Serina carboxipeptidasa	31	64
Actina	17	41
Ferritina	3	7
Familia de proteínas PRA1	7	19
Enolasa	3	3
β-amilasa	6	8
Factor de transcripcion bZIP	14	36
Glutaredoxina.	18	31

La comparación del número de miembros de la familia de genes identificados en el índice de genes de camote comparadas con las de *A. thaliana* indica que cubre cerca del 50% de la diversidad génica (277 miembros de las familia de genes en el índice de genes / 653 miembros encontrados en *A. thaliana*, 42%). Por lo tanto, esto nos muestra que el índice de genes contiene cerca de la mitad del transcriptoma.

Dentro de las familias de genes más representadas tenemos al intercambiador de sodio e hidrógeno, enolasa, y entre los genes con menor número de miembros representados tenemos al factor de transcripción MYB y bHLH. De todos modos, el índice de genes no puede cubrir todos los genes en este trabajo encontrados en *A. thaliana* ya que sólo se utilizaron cDNA de hojas, tallos y raíces, pero los demás tejidos, así como las otras etapas de desarrollo, no estaban cubiertas.

4.1.8 Anotación de secuencias

Las anotaciones de todos los *contigs* y *singletons* están disponibles en la página https://research.cip.cgiar.org/confluence/display/SPGI/Home. Esta página web también incluye el BLAST para realizar una búsqueda de las secuencias del índice de genes.

Según los resultados obtenidos del proceso de anotación de secuencias de las 66,418 secuencias, 39,299 (59%) tienen un *hit* (E-value 10^{-10}) con la base de datos de proteínas del UniRef100. De estos últimos, 24,657 son secuencias únicas. Se obtuvo un mayor número de *hits* para los *contigs* que para los *singletons* (Figura 19).

El porcentaje de identidad en nuestro índice de genes entre *contig* y *singleton* y de las proteínas de las bases de datos UniRef100 que han sido utilizados para la anotación de nuestro índice de genes está entre 24 y 100%, con un E-value en un rango de 10^{-10} a 10^{-180}.

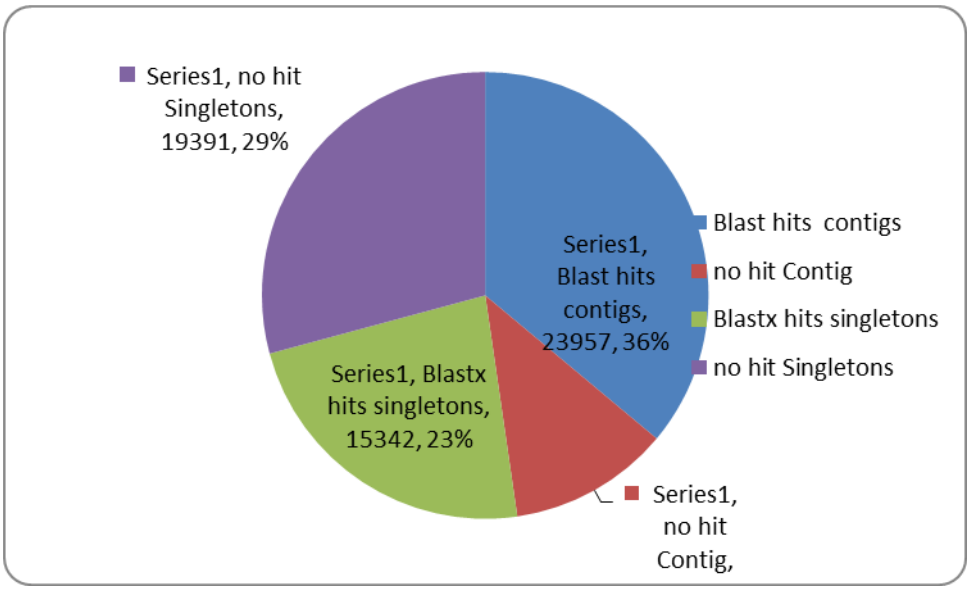

Figura 19. Distribución de los Blastx-*hits* y no-*hits* en los *contigs* y *singletons* del índice de genes anotadas con la base de datos del UniRef100.

Para demostrar que las secuencias que no tuvieron *hit* con la base de datos UniRef100 representan secuencias de genes y no artefactos, se evaluó la presencia de la cola Poly A y la presencia de ORF en los *contigs* y *singletons*. Se observó que de los 7,728 *contigs* que no tenían *hits* con la base UniRef100, 2,829 de estas secuencias contienen la cola PolyA, lo cual es un indicativo de que estas secuencias representen parte del gen transcrito. De 4,899 *contigs* restantes, 3,464 tienen un Open Reading Frame (ORF, contados a partir del codón de la posible metionina como codón de inicio hacia el codón *stop*), que codifican por lo menos 50 aminoácidos. Así, esto sugirió que prácticamente todos los *contigs* sin *hit* son derivados de genes que codifican proteínas. De la misma forma, se buscó la presencia de cola PolyA y ORF en los *singletons* que no tuvieron *hits*; así, de los 19,391 *singletons* que no tuvieron *hits* al compararlo con la base de datos del UniRef100, 6,775 tienen una cola poly-A. Solamente 635 *singletons* que no tienen *hits*, dentro de los cuales incluyen 47 ESTs del Genbank, no codifican ningún ORF que codificara más de 10 aminoácidos, mientras que el restante de los *singletons* sin *hits* fueron ORFs que tienen de 30 a 896 pb. La presencia de ORF en más del 90% de los *singletons* sin *hits* sugiere que estas secuencias también son derivadas de genes que codifican proteínas.

4.1.9 Asignación de GO terms

El programa BlastGO ha sido utilizado para atribuir términos de GO a los *contigs* y *singletons* del índice de genes. Los resultados nos muestran que del total de 66,418

secuencias, 38,763 tuvieron *hits* (23,641 *contigs* y 15,122 *singletons*), de estas últimas, a 30,693 (19,511 *contigs* y 11,182 *singletons*) se les asignaron el *GO terms* utilizando el Blast2GO. Para 27,655 secuencias no ha sido encontrado *hit* alguno con la base NCBI nr (secuencias de proteínas no redundates del NCBI). La representación gráfica que se presenta en la anotación corresponde al nivel 2 del *gene ontology* (Figura 20, Figura 21 y Figura 22).

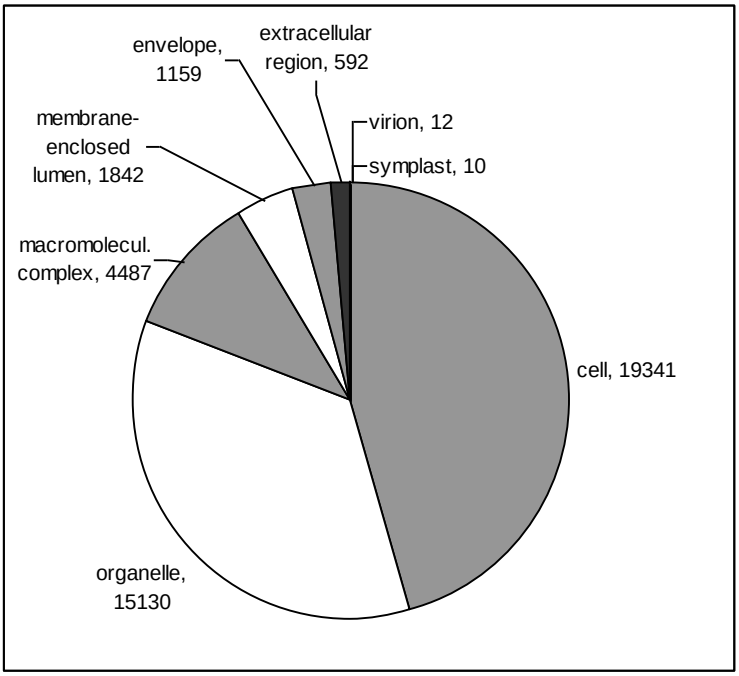

Figura 20. GO-anotación basado en la ubicación celular.

La anotación basada en la ubicación celular nos indica que se ha encontrado de manera decreciente un gran número de proteínas celulares, proteínas de organelos, complejos macromoleculares, proteínas de lumen, proteínas de envoltura celular además de proteínas de la región extracelular.

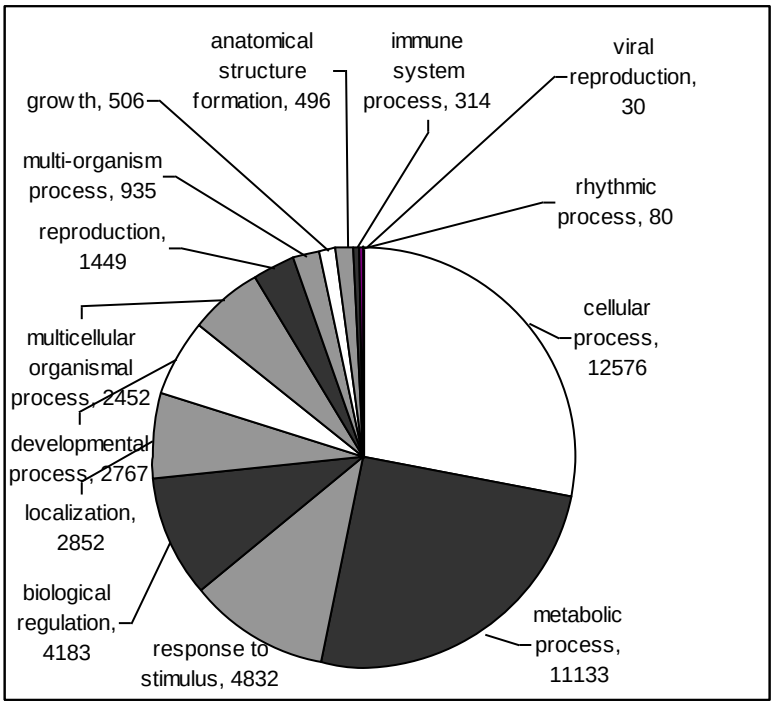

Figura 21. GO-anotación basado en el proceso biológico.

Las proteínas que se encontraron considerando el proceso biológico, se encuentran en un amplio rango de procesos, dentro de ellos tenemos a proteínas involucrados en procesos celulares, procesos metabólicos, respuesta ante estímulos, regulación biológica, localización, desarrollo, reproducción, crecimiento, proteínas que dan la estructura celular y otros (Figura 21).

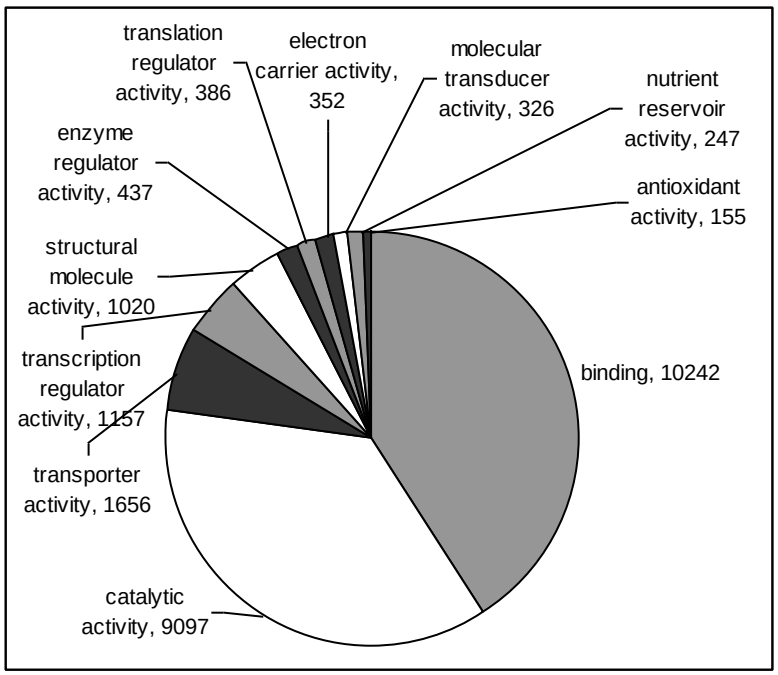

Figura 22. GO-anotación basado en la función molecular.

En la anotación según la función molecular que cumplen estas proteínas, tenemos un gran número de proteínas involucradas en interacciones moleculares (proteína – proteína, proteína ADN o ARN, proteína – membrana), proteínas con actividad catalítica, seguido de proteínas involucradas en la actividad transportadora, factores de transcripción, proteínas estructurales, enzimas reguladoras, proteínas reguladoras de la traducción, proteínas involucradas en el transporte de electrones, proteínas transductoras de señalización, proteínas de reserva, proteínas con actividad antioxidante (Figura 22).

No todos los *contigs* y *singletons* fueron derivados de camote, varias secuencias fueron claramente identificadas como secuencias virales (Tabla 4). Estas secuencias virales fueron derivadas de virus del *featheary mottle*, *chlorotic stunt* virus, *leaf curl* o virus con similaridad al virus A de la papa y *petunia vain clearing* virus que aparentemente han estado presentes en el experimento. También se han encontrado dos *singletons* que provenian de ESTs obtenidos del Genbank que representaban a transcriptos virales (Tabla 4).

Tabla 4. Secuencias virales contenidas en el índice de genes de camote.

Contig*/*singleton	**longitud**	**ID**	**E-value**	**Anotación**
03767	1851	Q59A02	0.0	Polyprotein (Fragment) n=1 Tax=Sweet potato feathery mottle virus RepID=Q59A02_9POTV
10592	1329	Q6XKE6	6e-49	Reverse transcriptase n=1 Tax=Petunia vein clearing virus isolate Hohn RepID=POLG_PVCV2
12069	526	O39734	3e-90	Polyprotein n=1 Tax=Sweet potato feathery mottle virus RepID=O39734_9POTV
12293	504	UPI0000163D32	6e-67	P3 protein n=1 Tax=Sweet potato feathery mottle virus RepID=UPI0000163D32
13375	686	O39734	2e-84	Polyprotein n=1 Tax=Sweet potato feathery mottle virus RepID=O39734_9POTV
23606	528	Q9QS58	1e-101	Coat protein AV1 n=1 Tax=Sweet potato leaf curl virus RepID=Q9QS58_9GEMI
27246	614	Q9QBT4	2e-67	Polyprotein n=1 Tax=Potato virus A RepID=Q9QBT4_PVMA
27646	725	UPI0000156FE9	1e-140	NIb protein n=1 Tax=Sweet potato feathery mottle virus RepID=UPI0000156FE9
29621	347	Q9QS57	2e-47	AC3 n=1 Tax=Sweet potato leaf curl virus RepID=Q9QS57_9GEMI
131893_1741_0589-I	221	UPI0000163D38	9e-36	coat protein n=1 Tax=Sweet potato feathery mottle virus RepID=UPI0000163D38
158451_1257_3402-I	250	B3Y549	4e-37	AV2 protein n=1 Tax=Sweet potato leaf curl virus RepID=B3Y549_9GEMI
DC882409.1	418	Q80MW4	2e-28	Polyprotein (Fragment) n=1 Tax=Sweet potato feathery mottle virus RepID=Q80MW4_9POTV
EE874850.1	669	Q8JJW9	1e-106	Polyprotein n=1 Tax=Sweet potato chlorotic stunt virus RepID=Q8JJW9_9CLOS
FRFM4LP02P1T45-II	318	UPI0000163D31	5e-34	HC-Pro protein n=1 Tax=Sweet potato feathery mottle virus RepID=UPI0000163D31
FRFM4LP02QRY4N-II	502	Q5GIU0	5e-45	Polyprotein (Fragment) n=1 Tax=Sweet potato feathery mottle virus RepID=Q5GIU0_9POTV
FRFM4LP02R1W0W-II	382	UPI0000163D34	4e-64	CI protein n=1 Tax=Sweet potato feathery mottle virus RepID=UPI0000163D34
FRFM4LP02RT0LA-II	328	C4N332	2e-36	Polyprotein n=1 Tax=Sweet potato feathery mottle virus RepID=C4N332_9POTV
FRFM4LP02RUC7U-II	451	O39734	4e-73	Polyprotein n=1 Tax=Sweet potato feathery mottle virus RepID=O39734_9POTV

Las secuencias obtenidas por la tecnología 454 han sido producidas con bibliotecas de cDNA derivadas de plantas expuestas a estrés hídrico. En consecuencia, esperamos que el

índice de genes presente un gran número de genes involucrados a estrés abiótico. Efectivamente, se ha encontrado un total de 199 *contigs* y *singletons* que tienen una GO-anotación "respuesta a estrés", 295 a "respuesta a estrés oxidativo", 209 a "respuesta a calor", 563 a "respuesta a estrés por salinidad", y 293 a "respuesta a carencia a agua". Ellos han sido identificados en plantas con estrés por sequía y que también han sido identificados en nuestro índice de genes (ver anexo 2).

Se realizó una comparación de otros trabajos de transcriptomas provenientes de secuenciamiento usando la tecnología 454. A pesar de ser difícil de comparar los resultados debido a que provienen de diferentes especies con diferente complejidad genética y han usado diferentes programas para el ensamblaje, sin embargo se puede observar que nuestro ensamblaje conduce a una reducción de la complejidad que va de 1.4 a 8.2 veces, siendo así nuestro trabajo más eficiente, comparado con otros trabajos. Considerando el número de *contigs* y *singletons*, NGen y Newbler son mucho más eficientes para reducir la complejidad comparado con TGICL, donde muchos *reads* se mantuvieron sin ensamblar (Tabla 5).

Tabla 5. Comparación proveniente de transcriptomas realizados con secuenciamiento 454.

Organismo	**Números de *reads***	**Tamaño de *reads* (pb)**	**ensamblaje**	**Número de *contigs***	**Número de *singletons***	**Reducción de complejidad**	**Referencia**
Ipomoea batatas	524,209 (+22094 Sanger *reads*)	244/313	NGen	31685	34733	8.2x	Schafleitner *et al* (submitido)
Melitaea cinxia	608,053	110	NGen	48,354	59,943	5.6x	Vera *et al*., 2008
Medicago truncatula	252,384	92	TGICL	33,865	150,734	1.4x	Cheung *et al*., 2006
Phytium ultimum	90,664	190	TGICL	11,155	24,352	2.6x	Cheung *et al*., 2008
Acropora millepora	623,267	216	Newbler/perl scripts	44,444	62,657	5.8x	Mayer *et al*., 2009
Eucalyptus grandis	1,024,251	102-210	Paracel Transcript Assembler, (Paracel Inc., Pasadena, CA)	71,384	No reportado		Novaes *et al*., 2009
Pinus taeda	586,732	306	Seqman Ngen (DNASTAR)	63,657	239,793	1.9x	Parchman *et al*., 2010

4.2 Obtención de los marcadores microsatélites

Un total de 1,621 motivos nuevos de SSR se identificaron en el índice de genes, esto incluye 936 di, 581 tri y 124 tetra nucleótidos con un mínimo de 7, 5, 5 unidades repetitivas respectivamente. El motivo repetitivo más frecuente observado fue el dinucleótidos TC, seguido de GA, el trinucleótidos TCT y el tetranucleótido AAAT. Se diseñó 223 SSRs de los cuales 195 amplificaron con el tamaño que se predijo, de estos 150 resultaron polimórficos y los 26 restantes fueron monomórficos, representando en total 729 alelos (4.1 alelos/locus) ver Tabla 6**.**

Tabla 6. Marcadores microsatélites en camote: Secuencias de los iniciadores y los motivos SSR identificados en las secuencias del índice de genes exitosamente amplificados

Primer	Forward primer	Reverse Primer	Motivo SSR	Ta	Tam	# alelos	Polimorfismo(p) Monomorfismo(m)	*Contig*/*singleton* ID
IbL1	TTTAGTCCCAGTATGACAATGC	GAAAGAACAGAGATCCTTCGAC	(AGA)6	60	156	2	p	17046
IbL10	TCGACGAGTGATCTGTAATCAT	GGAGAATCACACTCCATCATC	(CTC)6	60	151	2	p	21539
IbL11	GGGAGGGTTATAATTGTGGAC	GAAACAATTTCCTCCACCAC	(GCG)6	60	185	3	p	28809
IbL12	CGATGGACAGAACAACTACTTC	AAATAGCCTAAGACCAGAACCA	(AT)6	60	110	2	m	21624
IbL13	ACGAGAAGACCGAGAAACAG	CCTATACGATAAGAATGGCTGG	(GCG)6	60	125	2	m	06642
IbL14	TCTTCATCTCATCATCATCCTC	CTTCCTCGTACCTGCTAAGAAT	(TTG)6	60	103	3	p	09103
IbL15	AACCACCTCAGACTCCTAAAAA	TTTATCCACTCTCACAATCTGC	(AGG)6	60	152	3	m	17623
IbL16	GTCTTGCTGGATACGTAGAACA	GGGAGAAGTAAGAGAACCGATA	(TTC)7	60	161	9	p	21913
IbL18	GATTCAAATGGGGAGAATTG	GAGCTTCAATATCATTTTCTCTGA	(ATT)6	60	175	3	p	24109
IbL2	TCGTGCATCATGTTTAGGTC	CCAACATGAATATGTTCAAACG	(GAA)6	60	177	5	p	09150
IbL20	CTTCTACGACCAACTCTGATGA	AACTTTGTAGAGAAGATTCGGG	(AAG)6	60	156	3	p	23608
IbL21	GTTTTCTCGACGAAACCATT	TAACCAGAAGAGGAAACCCTAA	(GCC)6	60	127	n.d.	n.d.	23092
IbL22	GTCCGTCGATAAACACTACAAG	GTTCTGTGAGATGTGAATTTCC	(TCT)6	60	150	2	p	20973
IbL23	CCAGTCAGTCACATCTCTCAAG	GCCAGAATCGAAGACTATATCC	(GAT)6	60	177	5	p	13700
IbL24	AAAAATGTACTCGGCGGTGC	GGTGAGAACAAGACAAGCATC	(ACC)6	60	142	2	p	28849
IBL25	TTCTTCACTCCACGTCAGTAAC	TGGGTAGAGAAATGTGTCTTTG	(CTT)8	60	168	4	p	24401
IbL31	GTTCAAAACAGTTGGGTGAGA	AGGAACGCACAGAGACTTTATT	(ATC)6	60	159	2	p	16720
IbL32	GGGATGAAGGAGAGAATGAGTA	TTGAAAACCTAGAGAGAAAGGG	(TGA)8	60	158	9	p	1510/13512
IbL33	TCTTGTCGAGTACTTCAGATGG	TCTATGTCGGTCTGATGATGTT	(CAA)6	60	174	4	p	04750
IbL34	ATAAATCATATCCCATGGAGGG	TCATATGATCTCTATTGCGGAG	(TTA)6	60	169	3	p	003150_1297_3371
IbL35	GAGTTCTACATCGACACCGAC	GAACTCTACAGCTCATGACGAA	(TCC)6	60	145	5	p	011270_1347_2155
IbL36	CTTCATCTTGGGCTTCCGAT	CTAAGACCTCCAAGATCCGAC	(TTG)8	60	158	3	p	017734_1289_0657
IbL37	TCACACACTTCATATGCCTACA	GGTTCCCGATCTGTGATAATAA	(TCC)7	60	150	2	p	029308_1564_1963
IbL40	CTAATCAGGGAGGACTTCCTTT	GTTACTCATCATTGGGTGGTC	(TGC)6	60	158	4	p	043456_1128_3657
IBL41	CAGATAAAATATCGTCGTGGTC	TATTCCTATCCCTCGTTGATCT	(GAA)7	60	156	3	p	04047
IbL42	AAGCATCGAGTGCGTATCGT	TTAGCGACGGAGTTGTACTGT	(ACC)6	60	117	5	p	046515_1487_0122
IbL43	GATATCTTTTCCACCTCCAAGA	GTGATGAAGCTTTCAGAAGTGA	(GAA)7	60	138	2	p	21851
IbL44	GCTAGGAGTTGGGAAGTCAG	CCAACAGTCATGGCTTCTAATA	(GAA)6	60	155	4	p	29325
IbL46	CTGAAATTAGGGATTGAAGAGG	TCCAATCACTCCTTGTTTTCTC	(AGA)9	60	171	6	p	12029/110909_1786_0720
IbL49	GGCATTTGTAGTACTCCATCCT	AAGAAGAAGAGAACCGCCAT	(TCG)6-(GCC)5	60	163	2	m	165955_1573_2579
IbL6	AGAAGCTGGTGGCTCGTAAC	GATGTACTTTGGAATGCTGATG	(GTC)5	60	170	4	p	08403

IbL9	TGTACAGAATAACCATCCCACT	CAGCCATGACTGTGATAAACTT	(AGA)7	60	168	5	p	02234
IbO2	TGTGGATCTGTTCTTTGAACC	TTCCATGTGGAGTGTGAAGTAT	(CTAT)8	60	175	9	p	18128
IbO3	GATCCTCCAATCTGCACTCTA	TGAATGCGACACTAAAACACT	(GACA)5	60	157	4	p	4964/FRFM4LP02RDYSH
IbO4	AATTAGACAGTCGTTTTAGGGC	CGCAGAGTAGATCAAGCATAAG	(ATAC)5	60	112	3	p	24990
IbO5	CGCAAATGACTGATTCTGTACT	CCATTACACTTACAGCTGCTTG	(ACAT)5	60	161	2	p	12090/023644_1491_0354
IbO7	TGAGTGGAGGAACAATGTAAAG	CCAGCACAACAAATTGCTACTA	(AGCC)5	60	162	3	p	05712
IBS1	GATATTCTGTTTGCTGCTCTCA	CGTCATCATCATAATCTTCCTC	(TCC)7	60	134	5	p	03529
IBS10	ATCCATCTCATCCACAACAA	CAACTTAACGGACATACGTACTCT T	(TGCA)5-(ATCC)5	60	200	4	p	19622
IBS100	TGCTATAGTTACGTGGACGAAG	TTTAATGCTGATGTGGATGC	(AAT)7-(AAC)7	60	167	6	p	FRFM4LP02RC6AP
IBS103	CATTATCTCCATCAACTCCTCC	AGCCTGAGAAGAATGGTAATGT	(CT)10	60	120	4	p	FRFM4LP02S25FZ
IBS104	CCTATTGGTGTCCTTTGGATT	AAGAAGAAGAGGAAGAAGACGA	(ATC)6-(TCT)5	60	129	3	p	12241
IBS106	ACCAGAAAAGGAAAAGTAGCAG	TGCAAGAAGATTAGTGGACAAC	(AT)11	60	176	n.d.	n.d.	FRFM4LP02PNQ89
IBS107	GCAAGCTGTGGCATTAAATC	GAACCCTAAAAAGAAGAACCCT	(TTC)7	60	194	5	p	FRFM4LP02T1EVY
IBS108	CTCTTTATTCAAGTTTCCCCAC	CAGGTTGACATGTTTACTGAGC	(TCT)7	60	184	5	p	FRFM4LP02QV7ZQ
IBS109	GTCGCTATGTGTTGACTGTTG	TTATCCTCAACGTGCAAATG	(AATA)5	60	136	7	p	FRFM4LP02QEV9O
IBS11	ACCTTCATCCCATCTCTCTCTA	GGAGCTGGAGAAGCTTTAGAT	(TATC)11	60	128	7	p	FRFM4LP02SNK8Z
IBS110	AGACGGGAAGAATATCAATGAG	TTAGATCCATGTCCTGAGTTTG	(AAG)7	60	160	4	p	FRFM4LP02RKLTA
IBS111	TGGGATTGTGAACTTGGATAG	ATGATCCATCATCCATCACA	(TGA)9	60	104	2	m	FRFM4LP02SAMP2
IBS113	AGCACTAGACTACCTGAGGAGG	TAGCATGCTGCGTCTACAAT	(ATT)7	60	196	3	p	FRFM4LP02PT9US
IBS115	TATCATTCTACCATGTTGACGC	GATTCCTGGGAAATAATGAAGC	(TTA)8	60	163	2	p	02875
IBS117a	GATGGAGGAAGATATTTGGGT	CTTTCTTTCTTTCTGCTTCTGC	(CT)10	60	110	9	p	7126/FRFM4LP02SL80T
IBS118	GGTAGAAGACGATGACGAAGAC	AAAATTAGTCCAACTCCCACAC	(ACT)6	60	197	5	p	08316
IBS12	CAGTTATCAATTCCCACCTACC	TTGCTGTGTTATAGGCTTTGTC	(GA)10	60	170	8	p	28524
IBS120	GGAGCTGAAGCTCTCTTACAAT	TGAGGTAGATTGAGCTCTGAAA	(GAT)8	60	174	1	m	05224
IBS121	CTCTTGGTTCTGTGGACTGTAG	GTTAGGAAAGCAGAAGATAGGC	(CAC)8	60	178	2	m	02186
IBS122	CtAGCCATTGTGAATGTGTCTT	CTCGCTCAGATTTCTCTCTTTT	(TTTA)5	60	198	n.d.	n.d.	05391
IBS123	GGTGGGTATAGGAAGTCATCAT	CACAACTTGATCCTGTATCGTC	(ATG)8	60	112	1	m	06415
IBS126	TACACATTCGCCTACAGATTTC	GGGAAGACGTAGTAGGTCGTAG	(TC)14	60	126	2	p	18181
IBS129	ATATGTACCCACACCCCTATGT	GATACCAGGTGAGTAGGTGGTT	(AG)10	60	125	8	p	07188
IBS13	ATCCATGGTGGGGTACTAGACT	CCTATGTTTGTGAGAATGGACA	(TAT)5	60	195	9	p	16873
IBS132	GATCGAGAGCATCACAGATAAA	TTCACGTAAAGATATGGAACCC	(AAC)7	60	177	5	p	08707
IBS133	GATAATGGGCTTTGATGATAGG	GGTGGTGAACTAATTGCTCTCT	(ATC)8	60	156	3	p	21143
IBS134	CTTCAATCACCTGAAACTCTGA	AATATCGCTATGTTCTTGGGAC	(CT)11	55	143	6	p	00290
IBS137	TcAACAGACGTCTTCACTTACC	TCGATAGTATGATGTGAATCGC	(CTT)8	60	137	6	p	03051
IBS139	CTATGACACTTCTGAGAGGCAA	AGCCTTCTTGTTAGTTTCAAGC	(GA)7	60	196	9	p	04560
IBS14	ATCAGACATGCTTTTGTGAGAC	AGGGACTCACTTTCACTGCTAT	(TATG)7	60	175	4	p	26510
IBS140	AGCTGAGAAACGAGAGTAGGTC	CGCAAGATTACCCAACATAGTA	(ATCA)5	60	200	4	p	03513
IBS141	GAAGCAGTAGTTGTGTTGCTTT	CTCTATCTTTATCTCTTCCGGC	(CTTT)6	60	118	9	p	03406
IBS144	TCGAACGCTTTCTACACTCTT	CTGTGTTTATAGTCTCTGGCGA	(TTC)9	60	164	7	p	02819
IBS145	AAAGCCTAAAGCTCTCCCTC	ATGTTGAAAAGTCAGTGAGCAG	(TGGC)5	48	121	4	p	05627
IBS146	GCAAACCTCAAAAAGCGTAA	TAGAGGAATTGTAGGGAGTGGT	(GTCT)5	60	182	7	p	04503
IBS147	TGTGTACATGAGTTTGGTTGTG	GAAGTGCAACTAGGAAACATGA	(GCA)8	55	192	6	p	05491
IBS148	CAAGTACAGTGGCATATGGAAG	TTGCAATAGTAACCTCTGTGTG	(AATA)5	60	128	n.d.	n.d.	04558
IBS149	CCACCTCCTTAGGTATCAGACT	ACTACTAGCGCTGCAACCTTAT	(AGA)8	60	189	6 + 4	p	03864
IBS15	AACAACCTCAGAAACCTCAGTC	AGGTCCAGATTTGTTCTGATTC	(GTT)6-(GTT)5		151	5	p	26759
IBS150	AGTCCCTTGAATGTGTACCTCT	AGCTGCAATCATACAGTCAATC	(CT)13	60	196	2 + 4	p	09141
IBS151	ACCACCGTGAGAGGAATAAAAT	TCAGCAGTAGCTTTAACAAGGA	(TTCT)5	60	171	3	p	08069
IBS152	GCATCGGATTCATTATAGACAG	TCTCTCTAGCAACAGAAGAAACC	(GTT)8	60	100	6	p	084156_1589_1875
IBS153	GTGAAAATGTGCTCTTTAAGCC	ACACGTCTCATGTAAATGACCT	(TC)10	58	189	5	p	14758
IBS154	TGGTTGTATACCTCAACTCCAA	CTGCCTGGTATTAACACGAAA	(TTGT)5	60	196	n.d.	n.d.	06059/11244/FRFM4LP02S7Q30
IBS156	TTGATTCCACTATGACTTGAGC	ACACCAACCCTTATATGCTTTC	(AG)10	58	196	7 + 3	p	07833

IBS158	GAGTTGTCTGTTGTGCTTGAAT	GACTAAGCAGTGGAGAAGAAGAA	(GAT)7	55	150	2	m	11608
IBS159	GTTTGTTTCTGTGATTGGTCTG	GACACTTTTCCGTTAGACATGA	(CGAA)5	59	116	1	m	08174
IBS163	TCCACTGACAAGACTATCAAGC	GGGATTGAAGTTACGTTTTAGC	(GAAT)6	60	199	n.d.	n.d.	10890/21567
IBS164	CTCTTCGCCAACTCTAAATCTT	CGTACATTAAAAACAGACCGAC	(AAT)7	60	187	5	p	013594
IBS166	TCCGTCTTTCTTCTTCTTCTTC	ATACACTAACTGCATCCAAACG	(AG)10-(GAA)6	55	172	8	p	29919
IBS168	CTACACTAAAGAGCATACCCCC	TACAACTCACAACCCAGAAAAC	(TG)10	60	177	n.d.	n.d.	12903
IBS169	CGTACTATGTTTCCCCCATTAC	AATGCATCTACCCTCCTTACAC	(TTG)8	53	125	7	p	15078
IBS170	CAGGGCGAATAAGTCTAAAGAA	TCAACTGTGTGCATAGTTCAGT	(AG)11	60	149	5	p	16878
IBS172	CAAAAATACCCTTCCCTTGA	CTTCTAGCACCTATTTATCGGG	(AC)11	59	200	2	m	19267
IBS173	GTGAAAATGtgCTCTTTAAGCC	ACACGTCTCATGTAAATGACCT	(TC)10	60	196	4	p	14758
IBS174	AGAGAACAAAATCGGGAAGAAC	CGAAATAGAGATTGTAATGGGG	(AGA)7	60	143	4	p	23041
IBS175	GTGGAGAAGGGTAAAAGTGAGA	GCCTTTACTCCTATTCCTCATT	(TA)10	55	189	1	m	185458_1365_2719
IBS18	GCCAAGGATGAAGGATATAGAA	ACAAcCAAACTAGCTAAAAGCC	(ATG)7	60	175	6	p	00304
IBS180	AGTCAAGTCTGTCCACTCACTATT	ATCTGATGAGGAGGAAGAGAAA	(TCA)5	60	196	n.d.	n.d.	07665
IBS181	GAAGTTCAATAATTGGAGGAGC	CCAAGTTATATGCACAACCTGA	(TGG)5	60	180	n.d.	n.d.	25858
IBS182	AGAGTCATCGAACATGTCAAAC	ATCTTCGAGAGAGAGAAACCAA	(ACAT)5	57	147	2	m	FRFM4LP02TCF1K
IBS183	GGATTGTCAGACACAGAAGAGA	ATATGCAGCTCAGACATGAAAC	(AAC)5-(AAT)7	60	142	n.d.	n.d.	14078/26740
IBS184	CATTCAATTCCTTCCTTAGTCG	TTTAGTTACTGCGAAGAGACCC	(CT)6	60	113	1 + 13	p	FRFM4LP02Q24VZ
IBS185	CCTTGTTTATTCATCCAGCTCT	ATGTTGAAAAGTCAGTGAGCAG	(TGGC)6	60	128	5	p	27305
IBS186	CAGAAACAAGCAAAGATCTCAC	CTGTTGCTTCTCTTCTCCTTCT	(AAG)8	60	195	5	p	26702
IBS187	TTGTAGGAGTTCATCATGGCT	TTGATAGTGTTCTTCAAGGGAC	(AAAT)6	60	200	3	m	27453
IBS188	GTATTCGCTCCTAACTTTTTGG	GCTCGCCTATTTTCTCTCTCTA	(AAG)6	60	179	1	m	20602
IBS189	GACAGTTGTGGATGAAGGTTG	CCTCCTTGTTGATGAGTGAGTA	(AAAT)5	55	155	1	m	FRFM4LP02QVE09
IBS19	TCCTATGAGTGCCCTAAGAATC	CTCCTTCGTCTTCTTCTTCTTC	(GGA)8		151	6	p	07544
IBS194	CAACAAAAGTCTTGGACACAAC	ACAAATCTCTATCCTTCACGCT	(ATA)7	58	166	1	m	057337_1862_2769
IBS195	TTTACTGCGTACGTTGTTTGTC	ACTCCATCACTTTTACTCCCTG	(ATT)8	55	194	5	p	065898_1800_1212
IBS199	TAACTAGGTTGCAGTGGTTTGT	ATAGGTCCATATACAATGCCAG	(ACA)7	60	155	10	p	FRFM4LP02QW9DU
IBS20	CATCATCaCaGCCTACATAACC	ACCTTCAGATCAACAGTTTCCT	(AAT)7-(AAC)5	60	105	4	p	12788
IBS202	CTTGAAAAAGAGGCTTCTTAGC	CTCTGCTTTCTAAACTCGGATT	(TC)9	60	148	n.d.	n.d.	02777
IBS203	TTGTAAGATGATGAACAGAGGC	GAGTGCAAAAGGGAGCTTATAC	(AGA)8	57	179	1	m	Repeat-31127
IBS204	AGGAATTGAAGCTAAGAGCAAG	CTCTAGCAATTAAGCAAGCAAG	(AGAT)6	55	148	2	p	FRFM4LP02SJKHB
IBS21	ATCTTTGGGGGTTACTTCTCTT	GCTGCCAAATCACTATCAAAC	(TG)13	60	192	n.d.	n.d.	14216
IBS22	AATGATACCACAAGCAGAAGTG	GCTTCTTCATCTTCACTCAACTC	(TCT)8	60	200	5	p	09844
IBS23	GTTTCAACTCTCAACCAGTCAG	GGGGAAGAGAAGTTACAGAAAA	(TCT)5	60	187	4	p	14133
IBS24	AGTGCAACCATTGTAATAGCAG	TCCTTTCTTCATCATGCACTAC	(CTGC)7	60	147	6	p	13640
IBS25	AACATGATGAACACACCATCTC	GTTGCTGATGTTGAGGTAACTG	(CAG)7	60	112	n.d.	n.d.	15112
IBS28	ATATCTTCCAACAGTCTGCCTT	GCTTTCTGCTCTTCTTTCACTT	(TTC)8	60	176	3	p	28982
IBS3	CTTCTTTGATTGCTTCTAGCCT	ATTCATGATCTGATAGTGGTGG	(AAT)6	60	199	5	p	10385
IBS30	GGTCCTGTTAAAACAGCTCCTA	CCTGTATTTCCACAACCTACAA	(ATT)9	60	200	5	p	12752
IBS33	ATCTCTTCATACCAATCGGAAC	CAATGATAGCGGAGATTGAAG	(TCT)8	60	177	6	p	10297
IBS38	CAAATAGGAGGATACCTTAGCTG	ATAGGTTGTAGTAGGCGGAGAA	(AAAT)5	60	176	5	p	12498
IBS39	CGATGAGTAGTGAGGTGAATGT	GTCACATCTGAGAAGCATGAAC	(TTC)7	60	139	2	m	15396
IBS4	CTCTTCCTCCTCAGATTACCAC	CCAAGTTCCATATCACATCAAG	(ATC)7	60	152	5	p	13174
IBS40	AGTCTGAGCTCTAATGCTGTCA	AGCCATTGCTTGATACAAGTG	(GCT)5	60	100	3	p	23302
IBS44	TTAATACACATGCCTCTCCATC	TTCATTGTGACTGTGAGGAAG	(ATC)8	60	121	7	p	25774
IBS45	ATGAGGTTGAGGCTGAATTAGA	CACTTGAATTTTCCTTCTCTCC	(ATG)8	60	152	3	p	06772/08378
IBS46	GAGAGACAGAGATTGAAGGACC	CGGGAGTGTCAAAATCTACATA	(TC)12	60	117	n.d.	n.d.	08732
IBS47	GACATGTGAGCCTGTTCTTTTA	CACAGCAGGGAAGTATATGAAA	(CT)10	60	104	6	p	14914
IBS48	TGGAGTACTCGAGAGATGAGG	AAACACAGTGCTTACAGGGAAT	(CCTG)5	60	191	2	p	05831/FRFM4LP02R3R5T
IBS50	CTTTAAGGTTTAATCGAGAGGG	GGACACAAAAGTCTCATTTCAC	(GA)12	60	120	6	p	00560
IBS51	ATTCTTCACACCTTCCTTCAGT	ATAAGAGGAAGAGAAAGaGGGG	(AAG)8	60	166	5	p	07425

IBS52	AGTCGGAGAGAGATATGAGGAG	TAATTTCTCGCTTGCTATGC	(GA)14	60	118	3	m	FRFM4LP02QJR14
IBS53	GTTTCCACTGAAGCACAAGTT	CCACATTTTCTATTTGCCCT	(TATG)5	60	119	4	p	FRFM4LP02P69D8
IBS56	AGAAGGGGGAAAAAGTCTTAAC	TCTGATCTAGGCCCAATAAACT	(AATA)5	60	193	3	p	FRFM4LP02REMRA
IBS58	CCAAGAGTAACGATTACTGGCT	ACAGTTGTGACCATGTGAAAGT	(TCT)9	60	198	n.d.	n.d.	FRFM4LP02SCTT7
IBS61	GACAAACATCATCATCAGCATC	CAATCTCCTCTTTTCTACTCCC	(CAT)7	60	114	2	m	FRFM4LP02RABBE
IBS62	ATCGTTCACAGTGACTATCTCG	TGTCAATTGGACACCTCTGTAT	(CAGA)5	60	119	6	p	FRFM4LP02SCEVS
IBS64	GAGGGTTGAAGAGTTCATAAGG	AAAAGGGAGATCTTAGCTACCC	(CATA)5	60	165	7	p	FRFM4LP02QG5P9
IBS65	ATGTTTTGTCAAACTCTAGGGC	CTAACATGGGTATTTTGGGA	(AAAT)6	60	198	n.d.	n.d.	FRFM4LP02TRZN3
IBS68	CTCTCTCTCCTCCACAATTCTT	ATTTGGAGGTGAAGGTAGAGAA	(TC)11	60	162	2	m	FRFM4LP02QW8PI
IBS71	CTGTCTATGGCAAAATACTCCA	GGCTTGAGAGAGGTTACTGTTT	(AAAT)6	60	151	4	p	FRFM4LP02QCOA0
IBS72	CTACTCTCTGCTGGTTTATCCC	CTAGTGGTCTCTCTTCCTCCAC	(AC)10	60	188	7	p	FRFM4LP02PVTOZ
IBS73	TTCGCTAAACTCTAGTGTCGAA	TTCTGATGTTATCCAGAGATGG	(TATT)6	60	166	n.d.	n.d.	17889
IBS74	GGGGAAGAATCAATCATACTCT	GAAGAAAGTGGACATGGAAGAT	(GA)11	60	176	3	p	28125
IBS75	TATGTGTGATGATGATGAGCAG	ACTTGCTTGAGCTCTTCTTTG	(AAT)8	60	153	3	p	FRFM4LP02QOACU
IBS77	AGCCAAGAAGGCAACAATACT	AGTTTGTGCTACCCTCGTTTAT	(AG)10	60	107	n.d.	n.d.	FRFM4LP02RZ4TS
IBS78	TGTCTCTTTCTCCTCCAATTCT	AGTGCAGAAAGGATAGGATGTT	(TCT)7	60	168	5	p	FRFM4LP02RRPMA
IBS79	GTTGAGATCGAATCCTTGAGTT	CCAACGATCATACTCAATACCA	(CA)11	60	153	5	p	FRFM4LP02R46DJ
IBS80	GGGAGGTACCACTAATCTTGAA	CCTCCTTTTCCTCTTGAATATG	(AAG)7	60	101	8	p	FRFM4LP02TTTU8
IBS81	CAAAATCTCTAGCATACCCCTT	TCATCTCACTGACCAAGTATGC	(TGCA)5	60	100	2	m	FRFM4LP02QG4QK
IBS82	GACATAATTTGTGGGTTTAGGG	GAAATGGCAGAATGAGTAAGG	(TCA)7	60	137	6	p	27880
IBS84	CAAAGATGAAGCAAGTAAGCAG	ACTAATGTTGATCTACGGACCC	(AGAT)7	60	173	6	p	FRFM4LP02TGFYQ
IBS85	AACTACTCATGGGGAGAACAAC	CTAACGAAAGTTTGGACATCTG	(AC)12	60	174	6	p	FRFM4LP02TK43W
IBS86	AGAAACTGAAAACTAAGCTCGC	GCTATGCGTTTACAGAAACAAG	(CTT)7	60	159	7	p	FRFM4LP02QL07L
IBS87	CCCAAAACTACTCTTATAGCCG	AGATTCTGTATGGTTACCTGGG	(GAA)7	60	189	2	p	07237
IBS88	GATATACTGCTGCAATGGAGTG	CCAACTAAAAATAGAGGCAACC	(CT)12	60	164	4	p	FRFM4LP02R8CHO
IBS89	CTACGCAAAACAAAGCTATCAG	CAAATTCATCTCTTTCCCTCTC	(GA)12	60	113	4	p	FRFM4LP02P13XU
IBS9	ACCTAGTGCACACCATTGAGTA	GGACAACACACTTGGTTTTAAG	(TTTA)5	60	199	2	m	27731
IBS90	GTGGTGATGAGGAAACTGAAG	ACCTTCTCCAAGAACTCTTCCT	(GAT)9	60	111	3	p	FRFM4LP02R04AS
IBS91	GCAGCAGATGAACTATAAGCAG	CCTTCATATCACCCCTACAAGT	(ATGT)5	60	137	4	p	FRFM4LP02S2H1L
IBS94	TCACTTTCCTCCTTCTCTTCTC	GTTCGGCTATCTGTCAAAGAAC	(CT)11	60	134	n.d.	n.d.	FRFM4LP02R3VJY
IBS95	AGACGATGACTCCACTACCTTT	GGGCTTCTCTGCTATTATGAAC	(CCG)7	60	125	2	p	FRFM4LP02SQF7O
IBS96	ATGCCCAACTGTTATTACTAGG	TAGACTGTCTTTACGATGTGCC	(TTTA)5	60	104	2	p	FRFM4LP02SJVPE
IBS97	GTTACCAGGAATTACGAACGAT	CTCTCTACAAAAACTCACAGCG	(TG)10	60	180	6	p	FRFM4LP02RPWYJ
IBS98	TCTTCCTAGCTCTTGGACACAT	ATTAATATTGGTGGTGGTGGTG	(CTT)7	60	177	5	p	FRFM4LP02Q26YW
IBS99	GGTATGCCTCTCTTAGCTTCAT	AATCTGCAGCCATATTACTTCC	(GTAT)6	60	175	5	p	FRFM4LP02RHVOT
IbU1	GGCTTATGTAATAGATGCACCA	TGCTTCTATGCTCTTAAGGTTG	(TG)8	60	139	3	p	21614
IbU10	GTCTGCGTGTCCGTAGCATA	ATACGCACTTCATATACCGGC	(TC)7	60	116	3	p	23147
IbU11	AGCATGCGACTGATATTTAGG	ATATTCAAAACTGGCCGATG	(AT)7	60	112	5	p	28449
IbU12	GTGAGAGAAATCCCAAAGAGAG	ATTAGTCCTTGAAAGGCAGAGA	(TC)9	60	135	5	p	21926
IbU13	GCAACCAATCTACAGCAAACTA	CAGATAAAGTCCCCATTTCTTC	(AG)8	60	150	7	p	13836/068335_1865_2701
IbU14	GGCAAGGTTTCTCAAGTTGTTA	ACGAGATTACTTCAAATAGGGC	(TA)8	60	122	2	p	04543
IbU15a	TATCAGCAAGTTCCATATGAGG	CCTTCTTCTGCTTGTGGATATT	(AT)7	60	104	2	p	04757
IbU16	CGATCTTTCTGGAAATCTGACT	CAGAGTGATCAAGAATGCAACT	(TA)7	60	178	2	p	18214
IbU18	ATTGAGTCTCTCTCTGCCTTTC	AATACTCTGACAGCGATGATTG	(TC)8	60	175	2	p	08291
IbU20	GGAGAGCAAGTGGAGAAAGTAT	ACTCCTAGACCCACAATTGAAC	(TA)7	60	177	6	p	032980_1888_2125
IbU21	CATGTACTCATTGGAAAGGATG	CACATAGATAGCAGTTTGGCAT	(AT)7	60	153	2	m	035515_1209_2230
IbU22	CACAGCGTAGAACATGTGAAG	ACAATCTAGAATTCCCACCACT	(TC)9	60	121	2	p	23102
IbU23	CTCAGACACAGAGCACTCATCT	CCTCATCACTTCCACCTAACC	(GA)13	60	115	7	p	063266_1766_2698
IbU25	CTCGGATTGATCTATGTTGCT	TGGCCATTTAGTTTCACTCA	(TA)7	60	167	2	p	075729_1350_1303
IbU27	GAAAAGCTTGTATCTTGTGTGC	CTTGGTTTCTTTCTTTCTCACC	(TC)7	60	176	3	p	106339_1135_3328
IbU28	GAAAAGTCTGGATTGCAGTTTC	GACAGATTGGATCAAAAGGGTA	(TC)10	60	138	3	p	125004_1273_0206

IbU29	AGTGCTGGCAAGTTGTTTCT	CAGTGTTCTATTGCTCTTCACTTAC	(AT)7	60	177	5	p	12980
IbU30	CTATGAAATGTTGTTAGGGCAC	ACTGTACTGGTTGGTTGGTTAT	(AG)12	60	110	3	m	150849_1843_3592
IbU31	CCGCAGAAAAAGTTCAGATT	GCAACTTTTCTTCTTCCGTAAC	(CT)12	60	158	6	p	168238_1939_0637
IbU33	TTTGAAGAAGATGAGAGCGAC	TCAGAAAGACGATACACTAGAGAGA	(TC)8	60	153	7	p	169330_1735_1048
IbU34	CTTGTGTTTGTGTTTATGTCCC	CCCCTAACGCTTCTCATATTC	(GA)11	60	148	4	p	15802
IbU37	CCTTATTTGTTACTTGTTTGATGTG	ATGCTACATGCCTAAATGTCC	(AT)9	60	110	5	p	206194_1546_1332
IbU4	GGCTGGATTCTTCATATTTAGC	GCTTAATGGATCAGTAACACGA	(GA)9	60	174	6	p	02006
IbU5	TTTCAAGGATGTGGCTAATG	CCTGCTAATATAACCCAACCTC	(AT)8	60	180	5	p	17335
IbU6	GGGGTAGAGAGAAGAGAGTGAC	CCAGGTGAGAGTGTCTTTCAA	(TC)7	60	147	6	p	00990
IbU7	GAATCTCCTTTGCTGTTTGTCT	CACATAGGCACATACTCACCTT	(AT)7-(TAGC)4	60	150	6	p	21514
IbU9	AGCAAGAAAGGTCATCATCTG	ACGAAAGAGACGAACCCTAATA	(GA)7	60	128	2	p	28704

5 DISCUSIÓN

El camote [*Ipomoea batatas* (Linnaeus, 1753) Lamark] posee un alto contenido de betacaroteno, el cual es importante para cubrir la deficiencia de provitamina-A, y es altamente consumida en países en vías de desarrollo. El mejoramiento genético busca obtener cultivos altamente resistentes a plagas o a estres abiótico, para lo cual es necesario contar con información de genes involucrados. La generación de información genómica y genética es la base para empezar un programa de mejoramiento genético, donde el establecimiento de un índice de genes contribuiría enormemente en la búsqueda de genes específicos, asi como en el desarrollo de marcadores microsatélites.

Índice de genes

En la actualidad existe poca información a nivel genómico y genético disponible para camote en comparación con papa (http://www.ncbi.nlm.nih.gov/). Más aun, hay una mínima información respecto a secuencias de genes de camote, existiendo aproximadamente 9,166 genes ensamblados a partir de información sobre secuencias ESTs que se puedan encontrar en NCBI (http://plantta.jcvi.org/cgi-bin/plantta_release.pl, Childs *et al*., 2007). En el presente trabajo de tesis se obtuvo un índice de genes que aporta 17,000 genes más que los ya existentes, obtenidos a partir del transcriptoma de camote.

Síntesis y normalización de cDNA

Utilizando *SMART^{TM} Technology* para la síntesis de cDNA se obtuvieron secuencias completas de *mRNA* con presencia del extremo 5'. El tamaño del cDNA obtenido en este trabajo está por encima del tamaño promedio del cDNA de mamífero que es de 2.2 Kb (Lewin, 1980), pero muy semejante al obtenido en *Medicago truncatula* por Cheung *et al*. (2006), usando la misma tecnología. Otra ventaja de esta tecnología es que recupera gran parte del extremo 5' en el cDNA, lo que se pudo verificar por la presencia de un gran porcentaje de ORF hallados tanto en los *contigs* como en los *singletons*.

En este trabajo se decidió utilizar bibliotecas de cDNA normalizadas, lo que nos permitió obtener un transcriptoma representativo (Zhulidov *et al*., 2004). Es decir, la posibilidad de

obtener genes con un alto nivel de expresión, así como también, genes con un menor número de copias. A ello se suma el haber trabajado con hojas y tallos, y no tan sólo un órgano, para obtener el transcriptoma del camote, lo que sin duda aportó en el número diferencial de genes encontrados.

Una manera indirecta de evaluar la normalización de la biblioteca de cDNA es por la obtención de un gran número de secuencias únicas, las que pueden servir como indicativo de la diversidad de genes presentes en el transcriptoma. Cabe resaltar que la obtención de un gran número de secuencias únicas no sólo depende de una buena normalización de las bibliotecas de cDNA, sino también del tamaño y la cantidad de las secuencias, la calidad del secuenciamiento y por supuesto del ensamblaje realizado. Nuestro índice de genes posee 24,657 secuencias únicas, lo cual es un número considerable comparado con el número de secuencias halladas por otros autores como Cheung *et al*. (2006) y Parchman *et al*. (2010).

Secuenciamiento 454

A través de la técnica de secuenciamiento 454 se buscó la obtención de secuencias nuevas, lo cual nos permitió la obtención de una gran cantidad de información en *reads*. Ésta posee mayores ventajas comparativas con respecto a otras técnicas, como por ejemplo Illumina que produce fragmentos más pequeños y mayor cantidad de *reads*, pero sólo el 50% es utilizable, por lo que es más adecuado usarlo en otro tipo de trabajo (Harismendy *et al*., 2009). Para nuestros fines, debido a la naturaleza hexaploide y heterocigótica del camote, los tamaño de los *reads* generados por la tecnología 454 fueron adecuados (300 a 500 pb), a diferencia de las generadas por Illumina, que serían muy complicadas y muy tediosas para el ensamblaje *de novo*. El tamaño promedio de los *reads* obtenidos fue en el caso de tallos y hojas, mayor a los obtenidos por Cheung *et al*. (2006), Vera *et al*. (2008), Cheung *et al*. (2008), Meyer *et al*. (2009) y menor al obtenido por Parchman *et al*. (2010).

Ensamblaje

La gran variación alélica presente en el genoma de camote, por su naturaleza heterocigota y hexaploide, hace muy difícil el ensamblaje. No existen hasta la actualidad parámetros determinados para el ensamblaje de un transcriptoma de un organismo con un genoma tan complejo, por lo tanto fue necesario determinar los parámetros de ensamblaje empíricamente comparando los resultados obtenidos del ensamblaje variando uno de los parámetros importantes que es el MMP. Generalmente, parámetros muy exigentes son usados para el ensamblaje de EST o *reads*, así tenemos que Parchman *et al*. (2010) han usado un valor de 93% MMP para ensamblar *reads* provenientes de secuenciamiento 454 de diferentes *Pinus* ssp, y Vera *et al*. (2008), quienes han ensamblado *reads* de secuenciamiento 454 de un *pool* genético diverso de *Melitaea cinxia* (Lepidoptera), han usado 80% MMP. Considerando que el camote es heterocigoto, poliploide y además tiene una alta densidad de SNP, es de esperar que un ensamblaje muy exigente daría como resultado una alta redundancia de las secuencias. Si por el contrario se disminuye la exigencia, se incrementaría el riesgo de ensamblar *reads* no relacionados entre sí en *contigs* quimeras y por lo tanto fusionar *reads* derivados de miembros de familias de diferentes genes en un solo *contig*. No hay valores o parámetros de ensamblaje previos para plantas como el camote, por lo tanto, fue necesario determinar los parámetros óptimos de ensamblaje, para ello se realizaron algunos *test* como se mencionó en resultados, el Megablast, Blastclust y análisis de familia de genes, sin embargo, fue muy difícil poder determinar el parámetro óptimo del ensamblaje. Al final en el presente trabajo se consideró el 80 % de MMP como el mejor parámetro para el índice de genes.

Es importante indicar que los *contigs* (31,685) obtenidos en el presente trabajo constituyen el 93% (489,391) del total de *reads* (524,214) obtenidos en el secuenciamiento. Este resultado es semejante a los porcentajes del 91%, 90% y 88 % hallados por Vera *et al*. (2008), Meyer *et al*. (2010) y Novaes *et al*. (2008), respectivamente, a diferencia de sólo el 48% hallados por Parchman *et al*. (2010). Ello es un gran soporte para afirmar que nuestro ensamblaje fue bueno, ya que justamente existen muchos *reads* soportando un *contigs*.

Anotación

El 59% de los genes del índice fueron anotados con la base de datos UniRef100. Estos resultados son muy significativos comparados con los de Cheung *et al*. (2008), que fueron del 25%, y los de Parchman *et al*. (2010), que fueron de 32%. Este alto número de secuencias anotadas probablemente se debe al mayor tamaño de *contigs* obtenidos en nuestro ensamblaje. Además, cabe resaltar que esta anotación se realizó con un *E-value* más exigente, comparado con los de Vera *et al*. (2008) y Parchamn *et al*. (2010).

Es importante tomar en cuenta que el presente índice de genes contiene 24,657 secuencias únicas comparadas con las halladas por otros investigadores, que fueron de 17,000 (Parchman *et al*., 2010) y 9,000 (Vera *et al*., 2008).

GO terms

En nuestro índice de genes existe un gran número de secuencias anotadas con asignación de los *GO terms*, 30,763 secuencia que representan el 46% del total. Este valor es más alto al encontrado por Parchman *et al*. (2010) que fue de 31%. En la asignación de *GO terms* se puede ver que existe un mayor número de *contigs* (19,511) con *GO terms* en comparación a los *singletons* (11,182).

Genes virales

En el índice de genes se ha encontrado secuencias de origen viral, esto es un indicativo de que las plantas estuvieron infectadas, tanto las plantas que se usaron para nuestro trabajo como las plantas que se usaron para generar EST depositadas en el Genbank. En la Tabla 4 observamos la presencia de genes del *Virus Feathery Mottle* (03767, 12069, 12293, 13375, 27646, 29621, FRFM4LP02QRY4N-II, FRFM4LP02QRY4N-II, FRFM4LP02R1W0W-II, FRFM4LP02RT0LA-II, FRFM4LP02RUC7U-II) y *chlorotic stunt (*EE874850.1) los cuales son los virus que producen las mayores pérdidas en el rendimiento del camote (Loebestein y Thottappilly, 2009).

Genes de respuesta a sequía

Los genes anotados como "*response to water deprivation*" comprenden genes involucrados en la percepción del estrés hídrico y la señalización, tales como las quinasas (*contigs* 1152, 1635, 2923 y 5381) y las fosfatasas (*contigs* 3634, 18329, 19357 y 24076). Estos genes son requeridos para que la planta regule la respuesta hacia el estrés hídrico (Bartels y Sunkar, 2005). Otra categoría funcional de los genes relacionados a estrés hídrico es la de los genes relacionados al reconocimiento y el metabolismo de las hormonas, tales como ERD 15 (*contig* 1573), que codifican las enzimas de la síntesis de etileno (*contigs* 2742, 6752) que se encuentran inducidos en muchas plantas bajo estrés hídrico (Kiyosue *et al*., 1994). En el índice de genes obtenido en nuestro trabajo también está presente el ácido abcísico (*Singleton* DC880322.1). ABA es la hormona crucial involucrada en la señalización de estrés hídrico, la regulación de expresión de genes, comportamiento estomático y el desarrollo de la planta (Wasilewska *et al*., 2008). Se ha encontrado factores de transcripción como NAC-TF (*contig* 2046), factor de transcripción sensible a etileno (*contig* 9316) o DREB (*contigs* 1428, 9005, 18230 y 27987) que incrementan la expresión de genes inducidos a sequía (Agarwal *et al*., 2006). Otros genes típicos de respuesta al estrés hídrico son las proteínas de los canales de agua (*contigs* 1546 y 4647, 6627, 9834), que regulan el intercambio de agua entre los compartimentos celulares, y entre las células (Kaldenhoff *et al*., 2008). Finalmente, en el índice de genes aquí elaborado para el camote también encontramos proteínas que protegen a las células de efectos adversos de estrés a sequía, tales como las proteínas LEA (*Contig* 1352, 8988, 9100, 9111) (Hundertmark y Hincha, 2008). La amplia representación de genes de estrés hídrico aquí obtenidos hace de este índice de genes de camote una excelente fuente para aislar y caracterizar genes de tolerancia a estrés y abre la posibilidad de realizar ensayos entre especies y evaluar la tolerancia a estrés abiótico a nivel molecular.

Es importante considerar que el proceso de anotación es realizado por un análisis de similitud con genes depositados en las bases de datos. Existen muchas bases de datos tanto nucleotídicas como proteícas usadas para este proceso. Es importante resaltar que si bien es cierto nuestro sistema de anotación ha sido exigente (E- value = 10^{-10}), es importante considerar que existen errores que provienen de las bases de datos que se toman de referencia. Según un estudio de 37 familias de enzimas de Schnoes *et al*. (2009), existe una alta frecuencia de malas anotaciones en las bases de referencia, como el GenBank NR con

hasta 63% de anotación errónea en comparación a la base de datos del Swiss-Prot (cerca al 0% para muchas de las familias). En este trabajo hemos realizado la anotación basado en UniRef100, la cual está basada en todas las secuencias de UniProtKB, la cual posee tanto las secuencias revisadas manualmente como las generadas automáticamente por programas de computación. Si bien realizar la anotación con secuencias anotadas manualmente sería lo ideal, lamentablemente esta base de datos posee muy pocas secuencias (518,415), resultando insuficiente para usarla en procesos de anotación de transcriptomas o genomas.

Microsatélites

Como producto directamente utilizable a partir de índice de genes del camote generado como parte de esta tesis, está la obtención de un gran número de marcadores microsatélites, los cuales tienen un alto potencial para el mejoramiento genético de la especie. La contribución de este trabajo es el alto porcentaje de microsatélites amplificados (84%, 195 del total de 233). Lo que se pudo ver también es que los microsatélites son altamente replicables en el laboratorio y tienen un alto grado de polimorfismo (150 de 195), como los hallados en el *kit* de papa (Ghislain *et al*., 2009). Existen varios grupos de microsatélites que han sido diseñados y utilizados en camote (Jarret y Bowen, 1994; Buteler *et al*., 1999; Zhang *et al*., 2000; Hu *et al*., 2004; Arizio *et al*., 2008; Veasey *et al*., 2008), pasan a ser ahora apenas un pequeño porcentaje de los marcadores microsatélites para el camote.

6 CONCLUSIONES

1. Basado en un ensamblaje híbrido de las secuencias 454 y las la secuencias EST del GenBank, se ha establecido un índice de genes de *Ipomoea batatas* (L.) Lam. con 66,418 secuencias nucleotídicas, conformado por 31,685 *contigs* y 34,733 *singletons* tomando como base de datos UniRef100.

2. Debido a la naturaleza hexaploide y heterocigótica del camote, el índice de genes contiene una redundancia porque el proceso de ensamblaje no permite una diferenciación estricta entre las variantes alélicas de un gen y los genes duplicados.

3. Se ha anotado 39,299 secuencias que representa el 59 % (23,957 *contigs* y 15,342 *singletons*) y se determinó 24,657 secuencias únicas anotadas usando como base de datos a UniRef100. Basado en número de secuencias anotadas y secuencias únicas, la redundancia del índice de genes fue estimada a 1.6 (en promedio 1.6 secuencias del índice de genes por cada locus). Del mismo modo, se ha logrado anotar las secuencias utilizando el *GO terms* de 30,693 secuencias entre *contigs* y *singletons*.

4. Se ha identificado un gran número de secuencias candidatas para marcadores microsatélites y se han diseñado iniciadores para la amplificación de esas secuencias a partir del índice de genes de *Ipomoea batatas* (L.) Lam. Se diseñaron 233 pares de iniciadores, de los cuales 195 han amplificado exitosamente sólo un locus de microsatélites, 150 de esos loci fueron polimórficos en los 8 genotipos de camote seleccionados.

5. El índice de genes (http://www.cipotato.org/sweetpotato_gene_index) obtenido incrementa fuertemente la información genómica y permite el diseño de marcadores moleculares y otras herramientas para *Ipomoea batatas*.

7 REFERENCIAS BIBLIOGRAFICAS

1. Ahn P.M. 1993. Tropical soils and fertilizer use. Intermediate Tropical Agriculture Series. Longman Scientific and Technical; Essex; UK.

2. Altschul SF; Madden TL; Schaffer AA; Zhang J; Zhang Z; Miller W; Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acid Res 25: 3389-402.

3. Agarwal PK; Agarwal P; Reddy MK; Sopory SK. 2006. Role of DREB transcription factors in abiotic and biotic stress tolerance in plants. Plant Cell Rep 25: 1263–1274.

4. Arizio CM; Hompanera N; Suarez EY; Manifesto MM. 2008. Genotypic identification and diversity evaluation of a sweet potato (*Ipomoea batatas* (L). Lam) collection using microsatellites. Plant Gen Res 7: 135-138.

5. Ashburner M; Ball CA; Blake JA; Botstein D; Butler H; Cherry JM; Davis AP; Dolinski K; Dwight SS; Eppig JT; Harris MA; Hill DP; Issel-Tarver L; Kasarskis A; Lewis S; Matese JC; Richardson JE; Ringwald M; Rubin GM; Sherlock G. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25: 25-29.

6. Austin DF. 1988. The taxonomy; evolution and genetic diversity of sweetpotatoes and related wild species. In: P. Gregory (ed.). Exploration; maintenance; and utilization of sweetpotato genetic resources; pp. 27–60. CIP; Lima; Peru.

7. Barnes WM. 1994. PCR amplification of up to 35-kb DNA with high fidelity and high yield from lambda bacteriophage templates. Proc Natl Acad Sci USA 91: 2216-2220.

8. Bartels D; Sunkar R. 2005. Drought and Salt Tolerance in Plants. Critical Rev Plant Sci 24: 23-58.

9. Bovell-Benjamin AC. 2007. Sweet potato: a review of its past; present; and future role in human nutrition. Adv Food Nutr Res 52: 1-59.

10. Buteler MI; Jarret RL; LaBonte DR. 1999. Sequence characterization of microsatellites in diploid and polyploid *Ipomoea*. Theor. Appl. Gen 99: 123–132.

11. Buteler MI; Labonte DR; Jarret RL; Macchiavelli RE. 2002. Microsatellite-based paternity analysis in polyploid sweet- potato. J. Am. Soc. Hortic. Sci. 127: 392-396.

12. Cervantes-Flores JC; Yencho GC; Kriegner A; Pecota KV; Faulk MA; Mwanga ROM; Sosinski BR. 2008. Development of a genetic linkage map and identification of homologous linkage groups in sweetpotato using multiple-dose AFLP markers. Mol Breeding 21: 511–532.

13. Chevreux B; Pfisterer T; Drescher B; Driese AJ; Müller WEG; Wetter T; Suhai S.2004. Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. Genome Res 14: 1147-1159.

14. Childs KL; Hamilton JP; Zhu W; Ly E; Cheung F; Wu H; Rabinowicz PD; Town CD; Buell CR; Chan AP. 2007. The TIGR Plant Transcript Assemblies database. Nucleic Acids Res 35: D846-851.

15. Clarke AC. 2010. Origins and dispersal of the sweet potato and bottle gourd in Oceania: implications for prehistoric human mobility. PhD thesis; Massey University; Palmerston North; New Zealand.

16. Conesa A; Götz S; García-Gómez JM; Terol J; Talón M; Robles M. 2005. Blast2GO: a universal tool for annotation; visualization and analysis in functional genomics research. Bioinformatics 21: 3674–3676.

17. Da Maia LC; Palmieri DA; de Souza VQ; Kopp MM; de Carvalho FI; Costa de Oliveira A. 2008. SSR Locator: Tool for simple sequence repeat discovery integrated with primer design and PCR simulation. Int J Plant Genomics 2008, ID 412696.

18. DNASTAR. 2009. Manual SeqMan NGen™ Help For Macintosh® and Windows®; Version 2.0 DNASTAR; Inc.; Madison USA.

19. Doyle JJ; Doyle JL. 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochem Bulletin 19:11-15.

20. Droege M; Hill B. 2008. The Genome Sequencer FLX™ System—Longer *reads*, more applications, straight forward bioinformatics and more complete data sets. J Biotech 136:3-10.

21. FAOSTAT; http://faostat.fao.org/

22. Ghislain M; Nuñez J; Herrera MR; Pignataro J; Guzman F; Bonierbale M. Spooner DM. 2009. Robust and highly informative microsatellite-based genetic identity kit for potato. Mol. Breeding 23: 377-388.

23. Gianfranceschi L; Seglias N; Tarchini R; Komjanc M; Gessler C. 1998. Simple sequence repeats for the genetic analysis of apple. Theor Appl Genet 96: 1069–1076.

24. Gupta PK; Varshnet RK; Sharma PC; Ramesh B. 1999. Molecular markers and their applications in wheat breeding. Plant Breeding 118: 369-390.

25. Harismendy O; Ng PC; Strausberg RL; Wang X; Stockwell TB; Beeson KY; Schork NJ; Murray SS; Topol EJ; Levy S; Frazer KA. 2009. Evaluation of next generation sequencing platforms for population targeted sequencing studies. Gen Biol 10: R32.

26. He XQ; Liu QC; Ishiki K; Zhai H; Wang YP. 2006. Genetic Diversity and Genetic Relationships among Chinese Sweetpotato Landraces Revealed by RAPD and AFLP Markers. Breed Sci 56: 201-207.

27. Hu J; Nakatani M; Mizuno K; Fujimura T. 2004. Development and characterization of microsatellite markers in sweetpotato. Breeding Sci 54: 177-188.

28. Huang X; Madan A. 1999. CAP3: A DNA sequence assembly program. Genome Res 9: 868-877.

29. Hundertmark M; Hincha DK. 2008. LEA (late embryogenesis abundant) proteins and their encoding genes in Arabidopsis thaliana. BMC Genomics 9: 118

30. Islam S. 2006. Sweetpotato (*Ipomoea batatas* L.) leaf: its potential effect on human health and nutrition. J Food Sci 71: 13-21.

31. Jarret RL; Bowen N. 1994. Simple sequence repeats (SSRs) for sweet potato germoplasm characterization. Plant Gen Res Newsletter 100: 9-11.

32. Kaldenhoff R; Ribas-Carbo M; Flexas Sans J; Lovisolo C; Heckwolf M; Uehlein N. 2008. Aquaporins and plant water balance. Plant; Cell & Environment 31: 658-566

33. Kiyosue T; Yamaguchi-Shinozaki K; Shinozaki K. 1994. ERD15, a cDNA for a dehydration-induced gene from *Arabidopsis thaliana*. Plant Physiol 106: 1707.

34. Kriegner A; Cervantes JC; Burg K; Mwanga ROM; Zhang D. 2003. A genetic linkage map of sweetpotato (*Ipomoea batatas* (L.) Lam.) based on AFLP markers. Mol Breeding 11: 169-185.

35. Loebestein G; Thottappilly G. 2009. The Sweetpotato. Springer Verlag; New York; USA.

36. Low J; Kinyae P; Gichuki S; Anyango Oyunga M; Hagenimana V; Kabira J. 1997. Combating vitamin A deficiency through the use of sweetpotato. Results from Phase I of an action research project in South Nyanza; Kenya. International Potato Center, Lima, Peru.

37. Low JW; Arimond M; Osman N; Cunguara B; Zano F; Tschirley D. 2007. A food-based approach introducing orange-fleshed sweet potatoes increased vitamin A intake and serum retinol concentrations in young children in rural Mozambique. J Nutr 137: 1320-1327.

38. Margulies M; Egholm M; Altman WE; Attiya S; Bader JS; Bemben LA; Berka J; Braverman MS; Chen YJ; Chen Z; Dewell SB; Du L; Fierro JM; Gomes XV; Godwin BC; He W; Helgesen S; Ho CH; Irzyk GP; Jando SC; Alenquer ML; Jarvie TP; Jirage KB; Kim JB; Knight JR; Lanza JR; Leamon JH; Lefkowitz SM; Lei M; Li J; Lohman KL; Lu H; Makhijani VB; McDade KE; McKenna MP; Myers EW; Nickerson E; Nobile JR; Plant R; Puc BP; Ronan MT; Roth GT; Sarkis GJ; Simons JF; Simpson JW; Srinivasan M; Tartaro KR; Tomasz A; Vogt KA; Volkmer GA; Wang SH; Wang Y; Weiner MP; Yu P; Begley RF; Rothberg JM. 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature 437: 376-380.

39. Mason JB; Lotfi M; Dalmiya N; Sethuraman K; Deitchler M. 2001. The micronutrient report: current progress and trends in the control of vitamin A; iodine; and iron deficiencies. Ottawa; Canada: The Micronutrient Initiative/International Development Research Centre http://www.micronutrient.org/frame_HTML/resource_text/publications/mn_report.pdf.

40. NCBI; http://www.ncbi.nlm.nih.gov/sites/structure/?term=ipomoea

41. Papanicolaou A; Stierli R; Ffrench-Constant RH; Heckel DG. 2009. Next generation transcriptomes for next generation genomes using est2assembly. BMC Bioinformatics 10: 447.

42. Parchman TL; Geist KS; Grahnen JA; Benkman CW; Buerkle CA. 2010. Transcriptome sequencing in an ecologically important tree species: assembly, annotation, and marker discovery. BMC Genomics 11: 180.

43. Pertea G; Huang X; Liang F; Antonescu V; Sultana R; Karamycheva S; Lee Y; W*hit*e J; Cheung F; Parvizi B; Tsai J; Quackenbush J. 2003. TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. Bioinformatics 19: 651-652.

44. Powell W; Machray GC; Provan J. 1996. Polymorphism revealed by simple sequence repeats. Trends Plant Sci 7: 215- 222.

45. Quackenbush J; Cho J; Lee D; Liang F; Holt I; Karamycheva S; Parvizi B; Pertea G; Sultana R; White J. 2001. The TIGR Gene Indices: analysis of gene transcript sequences in highly sampled eukaryotic species. Nuc Ac Res 29: 159-164.

46. Röder MS; Körzun V; Wendehake K; Plaschke J; Tixier MH; Leroy P; Ganal MW. 1998. A microsatellite map of wheat. Genetics 149: 2007-2023.

47. Ronaghi M; Uhlen M; Nyren P. 1998. A sequencing method based on real-time pyrophosphate. Science 281: 363-365.

48. Shagin DA; Rebrikov DV; Kozhemyako VB; Altshuler IM; Shcheglov AS; Zhulidov PA; Bogdanova EA; Staroverov DB; Rasskazov VA; Lukyanov S. 2002. A novel method for SNP detection using a new duplex-specific nuclease from crab hepatopancreas. Genome Res 12: 1935-1942.

49. Soderlund C; Johnson E; Bomhoff M; Descour A. 2009. PAVE: Program for assembling and viewing ESTs. BMC Genomics 10: 400.

50. Suzek BE; Huang H; McGarvey P; Mazumder R; Wu C. 2007. UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics 23: 1282-1288.

51. CGIAR (Consultative Group on International Agricultural Research). Sweet Potato; http://www.cgiar.org/impact/research/sweetpotato.html

52. Van Jaarsveld PJ; Faber M; Tanumihardjo SA; Nestel P; Lombard CJ; Spinnler Benadé AJ. 2005. β-Carotene–rich orange-fleshed sweet potato improves the vitamin A status of primary school children assessed with the modified-relative-dose-response test. Am J Clin Nutr 81: 1080-1087.

53. Veasey EA; Borges A; Silva Rosa M; Queiroz Silva JR; De Andrade Bressan E; Peroni N. 2008. Genetic diversity in Brazilian sweet potato (*Ipomoea batatas* (L.) Lam.; Solanales; Convolvulaceae) landraces assessed with microsatellite markers. Genetics and Molecular Biology 31: 725-733.

54. Vera JC; Wheat CW; Fescemyer HW; Frilander MJ; Crawford D; Hanski I; Marden JH. 2008. Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. Mol Ecology 17: 1636-1647.

55. Wasilewska A; Vlad F; Sirichandra C; Redko Y; Jammes F; Valon C; Frei N; Leung J. 2008. An Update on Abscisic Acid Signaling in Plants and More. Molecular Plant 1: 198-217.

56. Woolfe JA. 1992. Sweetpotato: an untapped food resource. Cambridge Univ. Press and the International Potato Center (CIP). Cambridge; UK.

57. Ye XS; Al-Babili; Kloti A; Zhang J; Lucca P; Beyer P; Potrykus I. 2000. Engineering the provitamin A (ß-carotene) biosynthetic pathway into (carotenoid-free) rice endosperm. Science 287: 303-305.

58. Zerbino DR; Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18: 821-829.

59. Zhang DP; Carbajulca D; Ojeda L; Rossel G; Milla S; Herrera C; Ghislain. 2000. Microsatellite Analysis of Genetic Diversity in Sweetpotato Varieties from Latin America. CIP Program Report 1999-2000, Lima, Peru.

60. Zhang DP; Ghislain M; Huaman Z; Golmirzaie A; Hijmans R. 1998. RAPD variation in sweetpotato (*Ipomoea batatas* (L.) Lam) cultivars from South America and Papua New Guinea. Genetic Resources Crop Evol 45: 271-277.

61. Zhu YY; Machleder EM; Chenchik A; Li R; Siebert PD. 2001. Reverse transcriptase template switching: a SMART approach for full-length cDNA library construction. Biotechniques 30: 892-897.

62. Zhulidov PA; Bogdanova EA; Shcheglov AS; Vagner LL; Khaspekov GL; Kozhemyako VB; Matz MV; Meleshkevitch E; Moroz LL; Lukyanov SA; Shagin DA. 2004. Simple cDNA normalization using Kamchatka crab duplex-specific nuclease. Nucleic Acids Res 32: e37.

8 ANEXOS

Anexo 1: Extracción de RNA (TRIzol ® reagent)

1. Moler las muestras, de 100-200 mg de tejido con nitrógeno liquido para 1 ml de TRIZOL, transferir el tejido molido a tubos eppendorf de 2 mL.
2. Agregar 1 mL de TRIZOL y homogenizar manualmente y si se necesita usar vortex. Centrifugar a 13 200 rpm por 10 minutos a 4ºC, luego incubar el homogenizado por 5 minutos a temperatura ambiente.
3. Transferir el sobrenadante a tubos de 2 mL, agregar 250 µL de cloroformo por cada mililitro de TRIZOL.
4. Agitar los tubos vigorozamente con las manos por 15 segundos, luego incubar a temperatura ambiente por 2 a 3 minutos.
5. Centrifugar las muestras a 13 200 rpm por 15 minutos a 4ºC.
6. Transferir el sobrenadante a un tubo nuevo de 2 mL, aquí es importante evitar la contaminación con la interfase.
7. Precipitar el ARN, para esto agregar 650 µL de isopropanol por cada mililitro de TRIZOL.
8. Incubar las muestras a temperatura ambiente por 10 minutos o -20ºC por 3 horas.
9. Centrifugar a 10 000 rpm por 15 minutos a 4ºC.
10. Remover el sobrenadante con tips con filtro y suavemente secar el pellet.
11. Lavar el pellet con 1 mL ETOH 75%.
12. Secar la muestra a temperatura ambiente.
13. Resuspender el pellet en 100 µL de agua DEPC.

Anexo 2. Genes con una anotación GO "respuesta a carencia de agua"

***Contig*/ *Singleton* No.**	**Anotación**	
01573	UniRef100_A0EPI4	ERD15 n=1 Tax=Capsicum annuum RepID=A0EPI4_CAPAN
18426	UniRef100_A0EPI4	ERD15 n=1 Tax=Capsicum annuum RepID=A0EPI4_CAPAN
Repeat-30945	UniRef100_A2PZD7	Extensin like protein n=1 Tax=Ipomoea nil RepID=A2PZD7_IPONI
Repeat-31442	UniRef100_A2PZD7	Extensin like protein n=1 Tax=Ipomoea nil RepID=A2PZD7_IPONI
11553	UniRef100_A5B1L1	Putative uncharacterized protein n=1 Tax=Vitis vinifera RepID=A5B1L1_VITVI
14750	UniRef100_A5B802	Putative uncharacterized protein n=1 Tax=Vitis vinifera RepID=A5B802_VITVI
03080	UniRef100_A5BHW6	Whole genome shotgun sequence of line PN40024, scaffold_0.assembly12x n=1 Tax=Vitis vinifera RepID=A5BHW6_VITVI
05003	UniRef100_A5BHW6	Whole genome shotgun sequence of line PN40024, scaffold_0.assembly12x n=1 Tax=Vitis vinifera RepID=A5BHW6_VITVI
01411	UniRef100_A5BTH6	Putative uncharacterized protein n=1 Tax=Vitis vinifera RepID=A5BTH6_VITVI
02792	UniRef100_A5BTH6	Putative uncharacterized protein n=1 Tax=Vitis vinifera RepID=A5BTH6_VITVI
15350	UniRef100_A8MRX5	Uncharacterized protein At4g19230.2 n=1 Tax=Arabidopsis thaliana RepID=A8MRX5_ARATH
17550	UniRef100_A8SJR3	Putative aquaporin PIP2;3 n=1 Tax=Phaseolus vulgaris RepID=A8SJR3_PHAVU
08724	UniRef100_A9NUC8	Putative uncharacterized protein n=1 Tax=Picea sitchensis RepID=A9NUC8_PICSI
00745	UniRef100_A9P7V1	Predicted protein n=1 Tax=Populus trichocarpa RepID=A9P7V1_POPTR
07805	UniRef100_A9PHT9	Putative uncharacterized protein n=1 Tax=Populus trichocarpa RepID=A9PHT9_POPTR
17089	UniRef100_B0F815	Proton gradient regulation 5 n=2 Tax=Cucumis RepID=B0F815_CUCME
22459	UniRef100_B1B1T9	ABA 8-oxidase n=1 Tax=Lactuca sativa RepID=B1B1T9_LACSA
18370	UniRef100_B2M0U5	Aquaporin n=1 Tax=Manihot esculenta RepID=B2M0U5_MANES
25706	UniRef100_B4FNQ5	Histone H4 n=1 Tax=Zea mays RepID=B4FNQ5_MAIZE
05832	UniRef100_B5B3R3	Glucan water dikinase n=1 Tax=Solanum lycopersicum RepID=B5B3R3_SOLLC
09899	UniRef100_B6EBD6	Heat shock protein 90-2 n=1 Tax=Glycine max RepID=B6EBD6_SOYBN
21696	UniRef100_B6SME2	Ethylene-responsive transcription factor 4 n=1 Tax=Zea mays RepID=B6SME2_MAIZE
02477	UniRef100_B6T0P4	Histone H4 n=1 Tax=Zea mays RepID=B6T0P4_MAIZE
13973	UniRef100_B6T0U6	Histone H4 n=1 Tax=Zea mays RepID=B6T0U6_MAIZE
02878	UniRef100_B7FJL7	Putative uncharacterized protein n=1 Tax=Medicago truncatula RepID=B7FJL7_MEDTR

Repeat-31352	UniRef100_B7FM69	Putative uncharacterized protein n=1 Tax=Medicago truncatula RepID=B7FM69_MEDTR
28508	UniRef100_B8QJH2	Histidine kinase 3 (Fragment) n=1 Tax=Betula pendula RepID=B8QJH2_BETVE
08452	UniRef100_B9A8D8	Indole-3-acetic acid 14 transcription factor like n=1 Tax=Malus prunifolia RepID=B9A8D8_9ROSA
09834	UniRef100_B9GMB9	Aquaporin, MIP family, NIP subfamily n=1 Tax=Populus trichocarpa RepID=B9GMB9_POPTR
07936	UniRef100_B9GNZ1	Nbs-lrr resistance protein n=1 Tax=Populus trichocarpa RepID=B9GNZ1_POPTR
01861	UniRef100_B9GZJ6	Predicted protein n=1 Tax=Populus trichocarpa RepID=B9GZJ6_POPTR
28909	UniRef100_B9HTV3	Predicted protein n=1 Tax=Populus trichocarpa RepID=B9HTV3_POPTR
06009	UniRef100_B9I9S5	Multidrug resistance protein ABC transporter family n=1 Tax=Populus trichocarpa RepID=B9I9S5_POPTR
17897	UniRef100_B9IHK5	Predicted protein n=1 Tax=Populus trichocarpa RepID=B9IHK5_POPTR
16860	UniRef100_B9INA7	Predicted protein n=1 Tax=Populus trichocarpa RepID=B9INA7_POPTR
03678	UniRef100_B9N0P8	Predicted protein n=1 Tax=Populus trichocarpa RepID=B9N0P8_POPTR
03744	UniRef100_B9R985	Protein HVA22, putative n=1 Tax=Ricinus communis RepID=B9R985_RICCO
17941	UniRef100_B9RFD7	Epoxide hydrolase, putative n=1 Tax=Ricinus communis RepID=B9RFD7_RICCO
19357	UniRef100_B9RFF2	Protein phosphatase 2c, putative n=1 Tax=Ricinus communis RepID=B9RFF2_RICCO
29186	UniRef100_B9RI18	Acetylornithine aminotransferase, putative n=1 Tax=Ricinus communis RepID=B9RI18_RICCO
27783	UniRef100_B9RMQ6	Signal transducer, putative n=1 Tax=Ricinus communis RepID=B9RMQ6_RICCO
16516	UniRef100_B9RYP2	Transcription factor, putative n=1 Tax=Ricinus communis RepID=B9RYP2_RICCO
28904	UniRef100_B9RYP2	Transcription factor, putative n=1 Tax=Ricinus communis RepID=B9RYP2_RICCO
04282	UniRef100_B9RZ04	Nodulin-26, putative n=1 Tax=Ricinus communis RepID=B9RZ04_RICCO
06607	UniRef100_B9RZ04	Nodulin-26, putative n=1 Tax=Ricinus communis RepID=B9RZ04_RICCO
15443	UniRef100_B9RZ04	Nodulin-26, putative n=1 Tax=Ricinus communis RepID=B9RZ04_RICCO
01546	UniRef100_B9S0D7	Tonoplast intrinsic protein, putative n=1 Tax=Ricinus communis RepID=B9S0D7_RICCO
21859	UniRef100_B9S2R0	ATP binding protein, putative n=1 Tax=Ricinus communis RepID=B9S2R0_RICCO
00605	UniRef100_B9S366	Phosphorylase n=1 Tax=Ricinus communis RepID=B9S366_RICCO
02923	UniRef100_B9S768	Serine/threonine-protein kinase SAPK10, putative n=1 Tax=Ricinus communis RepID=B9S768_RICCO
26636	UniRef100_B9SER1	Oligopeptide transporter, putative n=1 Tax=Ricinus communis RepID=B9SER1_RICCO
04042	UniRef100_B9SI27	Epidermis-specific secreted glycoprotein EP1, putative n=1 Tax=Ricinus communis RepID=B9SI27_RICCO
30628	UniRef100_B9SRD2	F-box/leucine rich repeat protein, putative n=1 Tax=Ricinus communis RepID=B9SRD2_RICCO
02046	UniRef100_C0J1R5	NAC domain protein NAC4 n=1 Tax=Gossypium hirsutum RepID=C0J1R5_GOSHI
08846	UniRef100_C4P	Calcineurin B-like protein 01 n=1 Tax=Vitis vinifera

	7Y8	RepID=C4P7Y8_VITVI
18701	UniRef100_C4P7Y8	Calcineurin B-like protein 01 n=1 Tax=Vitis vinifera RepID=C4P7Y8_VITVI
22104	UniRef100_C4P7Y8	Calcineurin B-like protein 01 n=1 Tax=Vitis vinifera RepID=C4P7Y8_VITVI
18690	UniRef100_C6SY55	Putative uncharacterized protein n=1 Tax=Glycine max RepID=C6SY55_SOYBN
07554	UniRef100_D0EWD6	Heat shock protein 90 n=1 Tax=Ipomoea nil RepID=D0EWD6_IPONI
08691	UniRef100_D0EWD6	Heat shock protein 90 n=1 Tax=Ipomoea nil RepID=D0EWD6_IPONI
16629	UniRef100_D0EWD6	Heat shock protein 90 n=1 Tax=Ipomoea nil RepID=D0EWD6_IPONI
20653	UniRef100_D0EWD6	Heat shock protein 90 n=1 Tax=Ipomoea nil RepID=D0EWD6_IPONI
02570	UniRef100_D1H190	Whole genome shotgun sequence of line PN40024, scaffold_0.assembly12x (Fragment) n=1 Tax=Vitis vinifera RepID=D1H190_VITVI
12716	UniRef100_D1H4A9	Whole genome shotgun sequence of line PN40024, scaffold_4.assembly12x (Fragment) n=1 Tax=Vitis vinifera RepID=D1H4A9_VITVI
08080	UniRef100_D1H9U8	Whole genome shotgun sequence of line PN40024, scaffold_141.assembly12x (Fragment) n=1 Tax=Vitis vinifera RepID=D1H9U8_VITVI
15393	UniRef100_D1H9U8	Whole genome shotgun sequence of line PN40024, scaffold_141.assembly12x (Fragment) n=1 Tax=Vitis vinifera RepID=D1H9U8_VITVI
23257	UniRef100_D1H9U8	Whole genome shotgun sequence of line PN40024, scaffold_141.assembly12x (Fragment) n=1 Tax=Vitis vinifera RepID=D1H9U8_VITVI
26181	UniRef100_D1H9U8	Whole genome shotgun sequence of line PN40024, scaffold_141.assembly12x (Fragment) n=1 Tax=Vitis vinifera RepID=D1H9U8_VITVI
22002	UniRef100_D1HNU6	Whole genome shotgun sequence of line PN40024, scaffold_77.assembly12x (Fragment) n=1 Tax=Vitis vinifera RepID=D1HNU6_VITVI
05473	UniRef100_D1HPH8	Whole genome shotgun sequence of line PN40024, scaffold_98.assembly12x (Fragment) n=1 Tax=Vitis vinifera RepID=D1HPH8_VITVI
22811	UniRef100_D1HYK2	Whole genome shotgun sequence of line PN40024, scaffold_20.assembly12x (Fragment) n=1 Tax=Vitis vinifera RepID=D1HYK2_VITVI
04888	UniRef100_D1HYZ6	Whole genome shotgun sequence of line PN40024, scaffold_20.assembly12x (Fragment) n=2 Tax=Vitis vinifera RepID=D1HYZ6_VITVI
13705	UniRef100_D1HYZ6	Whole genome shotgun sequence of line PN40024, scaffold_20.assembly12x (Fragment) n=2 Tax=Vitis vinifera RepID=D1HYZ6_VITVI
04744	UniRef100_D1I2Q9	Whole genome shotgun sequence of line PN40024, scaffold_10.assembly12x (Fragment) n=1 Tax=Vitis vinifera RepID=D1I2Q9_VITVI
12462	UniRef100_D1I5Y7	Whole genome shotgun sequence of line PN40024, scaffold_59.assembly12x (Fragment) n=1 Tax=Vitis vinifera RepID=D1I5Y7_VITVI
06372	UniRef100_D1I5Y8	Whole genome shotgun sequence of line PN40024, scaffold_59.assembly12x (Fragment) n=1 Tax=Vitis vinifera RepID=D1I5Y8_VITVI
04108	UniRef100_D1I6U1	Whole genome shotgun sequence of line PN40024, scaffold_47.assembly12x (Fragment) n=2 Tax=Vitis vinifera

		RepID=D1I6U1_VITVI
06784	UniRef100_D1IBR0	Whole genome shotgun sequence of line PN40024, scaffold_3.assembly12x (Fragment) n=1 Tax=Vitis vinifera RepID=D1IBR0_VITVI
08953	UniRef100_D1IBR0	Whole genome shotgun sequence of line PN40024, scaffold_3.assembly12x (Fragment) n=1 Tax=Vitis vinifera RepID=D1IBR0_VITVI
18534	UniRef100_D1IBR0	Whole genome shotgun sequence of line PN40024, scaffold_3.assembly12x (Fragment) n=1 Tax=Vitis vinifera RepID=D1IBR0_VITVI
27659	UniRef100_D1IBR0	Whole genome shotgun sequence of line PN40024, scaffold_3.assembly12x (Fragment) n=1 Tax=Vitis vinifera RepID=D1IBR0_VITVI
08157	UniRef100_D1IC34	Whole genome shotgun sequence of line PN40024, scaffold_3.assembly12x (Fragment) n=1 Tax=Vitis vinifera RepID=D1IC34_VITVI
11641	UniRef100_D1IC34	Whole genome shotgun sequence of line PN40024, scaffold_3.assembly12x (Fragment) n=1 Tax=Vitis vinifera RepID=D1IC34_VITVI
19488	UniRef100_D1IDI0	Whole genome shotgun sequence of line PN40024, scaffold_19.assembly12x (Fragment) n=1 Tax=Vitis vinifera RepID=D1IDI0_VITVI
06395	UniRef100_D1IDJ9	Whole genome shotgun sequence of line PN40024, scaffold_19.assembly12x (Fragment) n=1 Tax=Vitis vinifera RepID=D1IDJ9_VITVI
26083	UniRef100_D1IDJ9	Whole genome shotgun sequence of line PN40024, scaffold_19.assembly12x (Fragment) n=1 Tax=Vitis vinifera RepID=D1IDJ9_VITVI
03855	UniRef100_D1IF21	Whole genome shotgun sequence of line PN40024, scaffold_26.assembly12x (Fragment) n=1 Tax=Vitis vinifera RepID=D1IF21_VITVI
07742	UniRef100_D1IJ12	Whole genome shotgun sequence of line PN40024, scaffold_51.assembly12x (Fragment) n=1 Tax=Vitis vinifera RepID=D1IJ12_VITVI
18131	UniRef100_D1IJ12	Whole genome shotgun sequence of line PN40024, scaffold_51.assembly12x (Fragment) n=1 Tax=Vitis vinifera RepID=D1IJ12_VITVI
01076	UniRef100_D1IKJ6	Whole genome shotgun sequence of line PN40024, scaffold_12.assembly12x (Fragment) n=1 Tax=Vitis vinifera RepID=D1IKJ6_VITVI
06584	UniRef100_D1IMD9	Whole genome shotgun sequence of line PN40024, scaffold_25.assembly12x (Fragment) n=1 Tax=Vitis vinifera RepID=D1IMD9_VITVI
30332	UniRef100_D1IMD9	Whole genome shotgun sequence of line PN40024, scaffold_25.assembly12x (Fragment) n=1 Tax=Vitis vinifera RepID=D1IMD9_VITVI
04465	UniRef100_D1IP77	Whole genome shotgun sequence of line PN40024, scaffold_81.assembly12x (Fragment) n=1 Tax=Vitis vinifera RepID=D1IP77_VITVI
06545	UniRef100_O04232	Cold-stress inducible protein n=1 Tax=Solanum tuberosum RepID=O04232_SOLTU
26364	UniRef100_O04232	Cold-stress inducible protein n=1 Tax=Solanum tuberosum RepID=O04232_SOLTU
25013	UniRef100_O09224	Water channel protein n=1 Tax=Nicotiana excelsior RepID=O09224_9SOLA
05015	UniRef100_O24448	Farnesyltransferase beta subunit (Fragment) n=1 Tax=Nicotiana glutinosa RepID=O24448_NICGU
01352	UniRef100_O65149	Late embryogenis abundant protein 5 n=1 Tax=Nicotiana tabacum RepID=O65149_TOBAC

08988	UniRef100_O65149	Late embryogenis abundant protein 5 n=1 Tax=Nicotiana tabacum RepID=O65149_TOBAC
09100	UniRef100_O65149	Late embryogenis abundant protein 5 n=1 Tax=Nicotiana tabacum RepID=O65149_TOBAC
09111	UniRef100_O65149	Late embryogenis abundant protein 5 n=1 Tax=Nicotiana tabacum RepID=O65149_TOBAC
15528	UniRef100_O65149	Late embryogenis abundant protein 5 n=1 Tax=Nicotiana tabacum RepID=O65149_TOBAC
21253	UniRef100_O65149	Late embryogenis abundant protein 5 n=1 Tax=Nicotiana tabacum RepID=O65149_TOBAC
25949	UniRef100_O65149	Late embryogenis abundant protein 5 n=1 Tax=Nicotiana tabacum RepID=O65149_TOBAC
04522	UniRef100_P27598	Alpha-1,4 glucan phosphorylase L isozyme, chloroplastic/amyloplastic n=1 Tax=Ipomoea batatas RepID=PHSL_IPOBA
25011	UniRef100_P27598	Alpha-1,4 glucan phosphorylase L isozyme, chloroplastic/amyloplastic n=1 Tax=Ipomoea batatas RepID=PHSL_IPOBA
06606	UniRef100_P32811	Alpha-glucan phosphorylase, H isozyme n=1 Tax=Solanum tuberosum RepID=PHSH_SOLTU
30741	UniRef100_P53535	Alpha-1,4 glucan phosphorylase L-2 isozyme, chloroplastic/amyloplastic n=1 Tax=Solanum tuberosum RepID=PHSL2_SOLTU
22071	UniRef100_P53537	Alpha-glucan phosphorylase, H isozyme n=1 Tax=Vicia faba RepID=PHSH_VICFA
03577	UniRef100_P93227	Protein farnesyltransferase/geranylgeranyltransferase type-1 subunit alpha n=1 Tax=Solanum lycopersicum RepID=FNTA_SOLLC
28997	UniRef100_Q0DC10	Os06g0498400 protein (Fragment) n=1 Tax=Oryza sativa Japonica Group RepID=Q0DC10_ORYSJ
06752	UniRef100_Q0H900	9-cis-epoxycarotenoid dioxygenase 3 n=1 Tax=Coffea canephora RepID=Q0H900_COFCA
26603	UniRef100_Q2HXJ3	Zeaxanthin epoxidase n=1 Tax=Chrysanthemum x morifolium RepID=Q2HXJ3_CHRMO
25131	UniRef100_Q2I6J7	Aquaporin (Fragment) n=1 Tax=Stevia rebaudiana RepID=Q2I6J7_STERE
22503	UniRef100_Q2XTE5	Hsp90-2-like n=1 Tax=Solanum tuberosum RepID=Q2XTE5_SOLTU
05536	UniRef100_Q309E7	Cryptochrome 2 n=1 Tax=Nicotiana sylvestris RepID=Q309E7_NICSY
05381	UniRef100_Q3HVN8	Serine/threonine protein kinase SAPK8-like protein n=1 Tax=Solanum tuberosum RepID=Q3HVN8_SOLTU
02742	UniRef100_Q3KRR3	9-cis-epoxycarotenoid dioxygenase 1 n=1 Tax=Cuscuta reflexa RepID=Q3KRR3_CUSRE
05420	UniRef100_Q40412	Zeaxanthin epoxidase, chloroplastic n=1 Tax=Nicotiana plumbaginifolia RepID=ABA2_NICPL
00344	UniRef100_Q41321	Protein induced upon tuberization n=1 Tax=Solanum demissum RepID=Q41321_SOLDE
00249	UniRef100_Q43797	Inorganic pyrophosphatase n=1 Tax=Nicotiana tabacum RepID=Q43797_TOBAC
16596	UniRef100_Q4SKJ3	Chromosome undetermined SCAF14565, whole genome shotgun sequence n=1 Tax=Tetraodon nigroviridis RepID=Q4SKJ3_TETNG
21776	UniRef100_Q5TJD4	Zinc finger DNA-binding protein n=1 Tax=Catharanthus roseus RepID=Q5TJD4_CATRO
18235	UniRef100_Q60E54	Os05g0389300 protein n=2 Tax=Oryza sativa RepID=Q60E54_ORYSJ
01635	UniRef100_Q66MH9	MAPKK n=1 Tax=Solanum lycopersicum RepID=Q66MH9_SOLLC
10396	UniRef100_Q66MH9	MAPKK n=1 Tax=Solanum lycopersicum RepID=Q66MH9_SOLLC
18832	UniRef100_Q66MH9	MAPKK n=1 Tax=Solanum lycopersicum RepID=Q66MH9_SOLLC
15061	UniRef100_Q66	Jasmonic acid 2 n=1 Tax=Solanum tuberosum RepID=Q66TP0_SOLTU

	TP0	
17836	UniRef100_Q6DLW2	Short-chain dehydrogenase/reductase n=1 Tax=Solanum tuberosum RepID=Q6DLW2_SOLTU
26062	UniRef100_Q6E4P4	Carotenoid cleavage dioxygenase 1B n=1 Tax=Solanum lycopersicum RepID=Q6E4P4_SOLLC
10806	UniRef100_Q6QPK1	AREB-like protein n=1 Tax=Solanum lycopersicum RepID=Q6QPK1_SOLLC
14144	UniRef100_Q6QPK1	AREB-like protein n=1 Tax=Solanum lycopersicum RepID=Q6QPK1_SOLLC
22999	UniRef100_Q6QPK1	AREB-like protein n=1 Tax=Solanum lycopersicum RepID=Q6QPK1_SOLLC
06720	UniRef100_Q6UJX4	Molecular chaperone Hsp90-1 n=1 Tax=Solanum lycopersicum RepID=Q6UJX4_SOLLC
00633	UniRef100_Q6UK17	HD-ZIP n=1 Tax=Capsicum annuum RepID=Q6UK17_CAPAN
05706	UniRef100_Q6UK17	HD-ZIP n=1 Tax=Capsicum annuum RepID=Q6UK17_CAPAN
25045	UniRef100_Q6XLQ1	Dehydrin-like protein n=1 Tax=Capsicum annuum RepID=Q6XLQ1_CAPAN
Repeat-31665	UniRef100_Q76MG1	Dehydrin (Fragment) n=1 Tax=Nicotiana tabacum RepID=Q76MG1_TOBAC
24686	UniRef100_Q7XAK5	Nitrate transporter n=2 Tax=Prunus persica RepID=Q7XAK5_PRUPE
00797	UniRef100_Q8H9B8	Low temperature and salt responsive protein n=1 Tax=Solanum tuberosum RepID=Q8H9B8_SOLTU
01530	UniRef100_Q8H9B8	Low temperature and salt responsive protein n=1 Tax=Solanum tuberosum RepID=Q8H9B8_SOLTU
01706	UniRef100_Q8H9B8	Low temperature and salt responsive protein n=1 Tax=Solanum tuberosum RepID=Q8H9B8_SOLTU
16436	UniRef100_Q8H9B8	Low temperature and salt responsive protein n=1 Tax=Solanum tuberosum RepID=Q8H9B8_SOLTU
25382	UniRef100_Q8H9B8	Low temperature and salt responsive protein n=1 Tax=Solanum tuberosum RepID=Q8H9B8_SOLTU
24478	UniRef100_Q8L8I6	Cytokinin receptor n=1 Tax=Catharanthus roseus RepID=Q8L8I6_CATRO
12358	UniRef100_Q8LJS1	No apical meristem-like protein n=1 Tax=Glycine max RepID=Q8LJS1_SOYBN
25963	UniRef100_Q8LKN6	Nam-like protein 18 n=1 Tax=Petunia x hybrida RepID=Q8LKN6_PETHY
03634	UniRef100_Q8RVG0	Protein phosphatase 2C n=1 Tax=Nicotiana tabacum RepID=Q8RVG0_TOBAC
18329	UniRef100_Q8RVG0	Protein phosphatase 2C n=1 Tax=Nicotiana tabacum RepID=Q8RVG0_TOBAC
24076	UniRef100_Q8RVG0	Protein phosphatase 2C n=1 Tax=Nicotiana tabacum RepID=Q8RVG0_TOBAC
29904	UniRef100_Q8RVG0	Protein phosphatase 2C n=1 Tax=Nicotiana tabacum RepID=Q8RVG0_TOBAC
23396	UniRef100_Q8VWX3	Hexokinase-related protein 1 n=1 Tax=Solanum tuberosum RepID=Q8VWX3_SOLTU
04422	UniRef100_Q8W1B0	Channel-like protein n=1 Tax=Petunia x hybrida RepID=Q8W1B0_PETHY
17264	UniRef100_Q94IB2	Phi-2 n=1 Tax=Nicotiana tabacum RepID=Q94IB2_TOBAC
03958	UniRef100_Q96578	Plasma membrane H+-ATPase n=1 Tax=Solanum lycopersicum RepID=Q96578_SOLLC
01891	UniRef100_Q9AWA5	Alpha-glucan water dikinase, chloroplastic n=1 Tax=Solanum tuberosum RepID=GWD1_SOLTU
06213	UniRef100_Q9AXF5	Starch phosphorylase (Fragment) n=1 Tax=Ipomoea batatas RepID=Q9AXF5_IPOBA

17758	UniRef100_Q9LKL9	Putative aquaporin TIP3 n=1 Tax=Vitis berlandieri x Vitis rupestris RepID=Q9LKL9_9MAGN
06634	UniRef100_Q9LKW3	Dehydration-induced protein ERD15 n=1 Tax=Solanum lycopersicum RepID=Q9LKW3_SOLLC
08262	UniRef100_Q9LKW3	Dehydration-induced protein ERD15 n=1 Tax=Solanum lycopersicum RepID=Q9LKW3_SOLLC
09201	UniRef100_Q9LKW3	Dehydration-induced protein ERD15 n=1 Tax=Solanum lycopersicum RepID=Q9LKW3_SOLLC
Repeat-31533	UniRef100_Q9LKW3	Dehydration-induced protein ERD15 n=1 Tax=Solanum lycopersicum RepID=Q9LKW3_SOLLC
01152	UniRef100_Q9M6Q9	MAP kinase kinase n=1 Tax=Nicotiana tabacum RepID=Q9M6Q9_TOBAC
04831	UniRef100_Q9M6Q9	MAP kinase kinase n=1 Tax=Nicotiana tabacum RepID=Q9M6Q9_TOBAC
08929	UniRef100_Q9SWA4	Histone H1C n=1 Tax=Nicotiana tabacum RepID=Q9SWA4_TOBAC
18413	UniRef100_Q9SWA4	Histone H1C n=1 Tax=Nicotiana tabacum RepID=Q9SWA4_TOBAC
09316	UniRef100_Q9SXS8	Ethylene-responsive transcription factor 3 n=1 Tax=Nicotiana tabacum RepID=ERF3_TOBAC
24788	UniRef100_Q9SXS8	Ethylene-responsive transcription factor 3 n=1 Tax=Nicotiana tabacum RepID=ERF3_TOBAC
Repeat-31305	UniRef100_Q9T0I8	Putative uncharacterized protein AT4g38800 n=1 Tax=Arabidopsis thaliana RepID=Q9T0I8_ARATH
04647	UniRef100_Q9XG70	Aquaglyceroporin (Tonoplast intrinsic protein (Tipa)) n=1 Tax=Nicotiana tabacum RepID=Q9XG70_TOBAC
06627	UniRef100_Q9XG70	Aquaglyceroporin (Tonoplast intrinsic protein (Tipa)) n=1 Tax=Nicotiana tabacum RepID=Q9XG70_TOBAC
09306	UniRef100_Q9XG70	Aquaglyceroporin (Tonoplast intrinsic protein (Tipa)) n=1 Tax=Nicotiana tabacum RepID=Q9XG70_TOBAC
08448	UniRef100_UPI00005A582F	PREDICTED: similar to germinal histone H4 gene n=1 Tax=Canis lupus familiaris RepID=UPI00005A582F
01559	UniRef100_UPI00005A5837	PREDICTED: similar to germinal histone H4 gene n=1 Tax=Canis lupus familiaris RepID=UPI00005A5837
01458	UniRef100_UPI0000DD8DD5	Os01g0835900 n=1 Tax=Oryza sativa Japonica Group RepID=UPI0000DD8DD5
18106	UniRef100_UPI0000DD8DD5	Os01g0835900 n=1 Tax=Oryza sativa Japonica Group RepID=UPI0000DD8DD5
20839	UniRef100_UPI00015CB289	PREDICTED: hypothetical protein n=1 Tax=Vitis vinifera RepID=UPI00015CB289
17572	UniRef100_UPI00015CC321	PREDICTED: hypothetical protein n=1 Tax=Vitis vinifera RepID=UPI00015CC321
04064	UniRef100_UPI0001982BBC	PREDICTED: hypothetical protein n=1 Tax=Vitis vinifera RepID=UPI0001982BBC
15606	UniRef100_UPI0001982DD4	PREDICTED: hypothetical protein isoform 1 n=1 Tax=Vitis vinifera RepID=UPI0001982DD4
26942	UniRef100_UPI0001982DD4	PREDICTED: hypothetical protein isoform 1 n=1 Tax=Vitis vinifera RepID=UPI0001982DD4
11299	UniRef100_UPI0001982E40	PREDICTED: hypothetical protein n=1 Tax=Vitis vinifera RepID=UPI0001982E40
18151	UniRef100_UPI0001982E40	PREDICTED: hypothetical protein n=1 Tax=Vitis vinifera RepID=UPI0001982E40
28412	UniRef100_UPI000198309F	PREDICTED: hypothetical protein n=1 Tax=Vitis vinifera RepID=UPI000198309F
28511	UniRef100_UPI0001983391	PREDICTED: hypothetical protein n=1 Tax=Vitis vinifera RepID=UPI0001983391
09647	UniRef100_UPI0001983901	PREDICTED: hypothetical protein n=1 Tax=Vitis vinifera RepID=UPI0001983901
18290	UniRef100_UPI	PREDICTED: hypothetical protein n=1 Tax=Vitis vinifera

	0001983901	RepID=UPI0001983901
28631	UniRef100_UPI0001983901	PREDICTED: hypothetical protein n=1 Tax=Vitis vinifera RepID=UPI0001983901
12344	UniRef100_UPI00019843FC	PREDICTED: hypothetical protein n=1 Tax=Vitis vinifera RepID=UPI00019843FC
29950	UniRef100_UPI00019844A0	PREDICTED: hypothetical protein n=1 Tax=Vitis vinifera RepID=UPI00019844A0
06202	UniRef100_UPI000198489D	PREDICTED: hypothetical protein n=1 Tax=Vitis vinifera RepID=UPI000198489D
22384	UniRef100_UPI000198489D	PREDICTED: hypothetical protein n=1 Tax=Vitis vinifera RepID=UPI000198489D
26396	UniRef100_UPI000198489D	PREDICTED: hypothetical protein n=1 Tax=Vitis vinifera RepID=UPI000198489D
12566	UniRef100_UPI0001984E0C	PREDICTED: hypothetical protein n=1 Tax=Vitis vinifera RepID=UPI0001984E0C
02290	UniRef100_UPI000198561F	PREDICTED: hypothetical protein n=1 Tax=Vitis vinifera RepID=UPI000198561F
043764_1320_2417	UniRef100_Q9LV58	Multiprotein-bridging factor 1c n=1 Tax=Arabidopsis thaliana RepID=MBF1C_ARATH
067802_1415_2604	UniRef100_UPI00019828F5	PREDICTED: similar to cytokinin receptor 1 n=1 Tax=Vitis vinifera RepID=UPI00019828F5
078572_1883_1217	UniRef100_Q3HLY8	U-box protein n=1 Tax=Capsicum annuum RepID=Q3HLY8_CAPAN
089201_1639_1731	UniRef100_B9SAP4	Multidrug resistance-associated protein 2, 6 (Mrp2, 6), abc-transoprter, putative n=1 Tax=Ricinus communis RepID=B9SAP4_RICCO
117926_1511_2637	UniRef100_UPI00019853F2	PREDICTED: hypothetical protein n=1 Tax=Vitis vinifera RepID=UPI00019853F2
159761_1683_0971	UniRef100_Q8H9B8	Low temperature and salt responsive protein n=1 Tax=Solanum tuberosum RepID=Q8H9B8_SOLTU
181054_1592_2522	UniRef100_Q8H9B8	Low temperature and salt responsive protein n=1 Tax=Solanum tuberosum RepID=Q8H9B8_SOLTU
193422_1881_2909	UniRef100_O65150	Osmotic stress-induced zinc-finger protein n=1 Tax=Nicotiana tabacum RepID=O65150_TOBAC
196206_1952_0810	UniRef100_Q2HXJ3	Zeaxanthin epoxidase n=1 Tax=Chrysanthemum x morifolium RepID=Q2HXJ3_CHRMO
CB330711.1	UniRef100_B6T0P4	Histone H4 n=1 Tax=Zea mays RepID=B6T0P4_MAIZE
CO499915.1	UniRef100_B9HLC2	Predicted protein (Fragment) n=1 Tax=Populus trichocarpa RepID=B9HLC2_POPTR
CO500785.1	UniRef100_B4FEB2	Putative uncharacterized protein n=1 Tax=Zea mays RepID=B4FEB2_MAIZE
DC880322.1	UniRef100_Q3HNF4	ABA 8'-hydroxylase CYP707A1 n=1 Tax=Solanum tuberosum RepID=Q3HNF4_SOLTU
DC880794.1	UniRef100_O24023	9-cis-epoxycarotenoid dioxygenase n=1 Tax=Solanum lycopersicum RepID=O24023_SOLLC
DC881247.1	UniRef100_O65149	Late embryogenis abundant protein 5 n=1 Tax=Nicotiana tabacum RepID=O65149_TOBAC
DV034738.1	UniRef100_Q0Q2I2	Aquaporin n=1 Tax=Tamarix sp. ZDY-001908 RepID=Q0Q2I2_9CARY
DV035037.1	UniRef100_P27598	Alpha-1,4 glucan phosphorylase L isozyme, chloroplastic/amyloplastic n=1 Tax=Ipomoea batatas RepID=PHSL_IPOBA
DV035120.1	UniRef100_P27598	Alpha-1,4 glucan phosphorylase L isozyme, chloroplastic/amyloplastic n=1 Tax=Ipomoea batatas RepID=PHSL_IPOBA
DV035186.1	UniRef100_C4P7W9	CBL-interacting protein kinase 20 n=1 Tax=Vitis vinifera RepID=C4P7W9_VITVI
DV035489.1	UniRef100_Q41321	Protein induced upon tuberization n=1 Tax=Solanum demissum RepID=Q41321_SOLDE
DV035491.1	UniRef100_B9SDI6	Spotted leaf protein, putative n=1 Tax=Ricinus communis RepID=B9SDI6_RICCO

DV035796.1	UniRef100_A2PZD7	Extensin like protein n=1 Tax=Ipomoea nil RepID=A2PZD7_IPONI
DV035870.1	UniRef100_A2PZD7	Extensin like protein n=1 Tax=Ipomoea nil RepID=A2PZD7_IPONI
DV035885.1	UniRef100_Q6UJX5	Molecular chaperone Hsp90-2 n=1 Tax=Nicotiana benthamiana RepID=Q6UJX5_NICBE
DV036020.1	UniRef100_A2PZD7	Extensin like protein n=1 Tax=Ipomoea nil RepID=A2PZD7_IPONI
DV036047.1	UniRef100_O81295	AT4g02380 protein n=1 Tax=Arabidopsis thaliana RepID=O81295_ARATH
DV036462.1	UniRef100_UPI00015CC229	PREDICTED: hypothetical protein n=1 Tax=Vitis vinifera RepID=UPI00015CC229
DV036496.1	UniRef100_D1IDJ9	Whole genome shotgun sequence of line PN40024, scaffold_19.assembly12x (Fragment) n=1 Tax=Vitis vinifera RepID=D1IDJ9_VITVI
DV036624.1	UniRef100_O24049	MipC n=1 Tax=Mesembryanthemum crystallinum RepID=O24049_MESCR
DV036626.1	UniRef100_B9IIZ5	Predicted protein n=1 Tax=Populus trichocarpa RepID=B9IIZ5_POPTR
DV037017.1	UniRef100_A2PZD7	Extensin like protein n=1 Tax=Ipomoea nil RepID=A2PZD7_IPONI
DV037347.1	UniRef100_A2PZD7	Extensin like protein n=1 Tax=Ipomoea nil RepID=A2PZD7_IPONI
DV037526.1	UniRef100_A2PZD7	Extensin like protein n=1 Tax=Ipomoea nil RepID=A2PZD7_IPONI
DV037534.1	UniRef100_Q6UJX5	Molecular chaperone Hsp90-2 n=1 Tax=Nicotiana benthamiana RepID=Q6UJX5_NICBE
DV037709.1	UniRef100_A2PZD7	Extensin like protein n=1 Tax=Ipomoea nil RepID=A2PZD7_IPONI
DV037829.1	UniRef100_A2PZD7	Extensin like protein n=1 Tax=Ipomoea nil RepID=A2PZD7_IPONI
EE877746.1	UniRef100_B9IKC3	Predicted protein n=1 Tax=Populus trichocarpa RepID=B9IKC3_POPTR
EE880224.1	UniRef100_B9IK05	NAC domain protein, IPR003441 n=1 Tax=Populus trichocarpa RepID=B9IK05_POPTR
EE882516.1	UniRef100_C0M519	CBL-interacting protein kinase 24 n=1 Tax=Populus euphratica RepID=C0M519_POPEU
FRFM4LP02P2G7N	UniRef100_P35016	Endoplasmin homolog n=1 Tax=Catharanthus roseus RepID=ENPL_CATRO
FRFM4LP02P7DTD	UniRef100_B9RZA9	Putative uncharacterized protein n=1 Tax=Ricinus communis RepID=B9RZA9_RICCO
FRFM4LP02P9TYL	UniRef100_Q6UN44	Phosphorylase (Fragment) n=1 Tax=Triticum aestivum RepID=Q6UN44_WHEAT
FRFM4LP02PGL7J	UniRef100_Q0ZCE9	Putative auxin-resistance protein (Fragment) n=1 Tax=Populus trichocarpa RepID=Q0ZCE9_POPTR
FRFM4LP02PL3F7		
FRFM4LP02PUJG7	UniRef100_C1IHU2	Histidine kinase 3B n=1 Tax=Populus trichocarpa RepID=C1IHU2_POPTR
FRFM4LP02PVTDM	UniRef100_UPI0001984707	PREDICTED: hypothetical protein n=1 Tax=Vitis vinifera RepID=UPI0001984707
FRFM4LP02PXKJ2	UniRef100_B6DXL4	NAC domain protein n=1 Tax=Malus x domestica RepID=B6DXL4_MALDO
FRFM4LP02PXMMW	UniRef100_Q40480	C-7 protein n=1 Tax=Nicotiana tabacum RepID=Q40480_TOBAC
FRFM4LP02PY8CX	UniRef100_Q2XTE5	Hsp90-2-like n=1 Tax=Solanum tuberosum RepID=Q2XTE5_SOLTU
FRFM4LP02Q1059	UniRef100_Q41413	Epoxide hydrolase n=1 Tax=Solanum tuberosum RepID=Q41413_SOLTU

FRFM4LP02Q787N	UniRef100_O24448	Farnesyltransferase beta subunit (Fragment) n=1 Tax=Nicotiana glutinosa RepID=O24448_NICGU
FRFM4LP02Q8HKP	UniRef100_Q1XHJ4	9-cis-epoxycarotenoid dioxygenase (Fragment) n=1 Tax=Citrus limon RepID=Q1XHJ4_CITLI
FRFM4LP02QA2BI	UniRef100_B9SZ64	Ring finger protein, putative n=1 Tax=Ricinus communis RepID=B9SZ64_RICCO
FRFM4LP02QEI00	UniRef100_B9HSM5	Predicted protein n=1 Tax=Populus trichocarpa RepID=B9HSM5_POPTR
FRFM4LP02QJO03	UniRef100_D0EWD6	Heat shock protein 90 n=1 Tax=Ipomoea nil RepID=D0EWD6_IPONI
FRFM4LP02QL2Q4	UniRef100_Q9LKW3	Dehydration-induced protein ERD15 n=1 Tax=Solanum lycopersicum RepID=Q9LKW3_SOLLC
FRFM4LP02QMZT9	UniRef100_O65150	Osmotic stress-induced zinc-finger protein n=1 Tax=Nicotiana tabacum RepID=O65150_TOBAC
FRFM4LP02QP1ZK	UniRef100_O65149	Late embryogenis abundant protein 5 n=1 Tax=Nicotiana tabacum RepID=O65149_TOBAC
FRFM4LP02QPP6N	UniRef100_B8QJH2	Histidine kinase 3 (Fragment) n=1 Tax=Betula pendula RepID=B8QJH2_BETVE
FRFM4LP02QQNUU	UniRef100_Q39193	Serine/threonine-protein kinase SRK2I n=1 Tax=Arabidopsis thaliana RepID=SRK2I_ARATH
FRFM4LP02QR0XC	UniRef100_B9I9S5	Multidrug resistance protein ABC transporter family n=1 Tax=Populus trichocarpa RepID=B9I9S5_POPTR
FRFM4LP02QWO7A	UniRef100_B6EBD6	Heat shock protein 90-2 n=1 Tax=Glycine max RepID=B6EBD6_SOYBN
FRFM4LP02R07YH	UniRef100_D1I2A7	Whole genome shotgun sequence of line PN40024, scaffold_10.assembly12x (Fragment) n=1 Tax=Vitis vinifera RepID=D1I2A7_VITVI
FRFM4LP02R38IY	UniRef100_A5BHW6	Whole genome shotgun sequence of line PN40024, scaffold_0.assembly12x n=1 Tax=Vitis vinifera RepID=A5BHW6_VITVI
FRFM4LP02R3AS7	UniRef100_A5ACW5	Putative uncharacterized protein n=1 Tax=Vitis vinifera RepID=A5ACW5_VITVI
FRFM4LP02RFVJ6	UniRef100_P53535	Alpha-1,4 glucan phosphorylase L-2 isozyme, chloroplastic/amyloplastic n=1 Tax=Solanum tuberosum RepID=PHSL2_SOLTU
FRFM4LP02RI1RQ	UniRef100_B9H0P2	Predicted protein (Fragment) n=1 Tax=Populus trichocarpa RepID=B9H0P2_POPTR
FRFM4LP02RL0SS	UniRef100_B9GL23	Predicted protein n=1 Tax=Populus trichocarpa RepID=B9GL23_POPTR
FRFM4LP02ROVAR	UniRef100_Q6UJX5	Molecular chaperone Hsp90-2 n=1 Tax=Nicotiana benthamiana RepID=Q6UJX5_NICBE
FRFM4LP02RQBAV	UniRef100_UPI00019853F2	PREDICTED: hypothetical protein n=1 Tax=Vitis vinifera RepID=UPI00019853F2
FRFM4LP02RRPOW	UniRef100_UPI00015CBF51	PREDICTED: hypothetical protein n=1 Tax=Vitis vinifera RepID=UPI00015CBF51
FRFM4LP02RWO0E	UniRef100_Q6UJX5	Molecular chaperone Hsp90-2 n=1 Tax=Nicotiana benthamiana RepID=Q6UJX5_NICBE
FRFM4LP02S2BO1	UniRef100_D1IC34	Whole genome shotgun sequence of line PN40024, scaffold_3.assembly12x (Fragment) n=1 Tax=Vitis vinifera RepID=D1IC34_VITVI
FRFM4LP02SBI24	UniRef100_UPI00019828A8	PREDICTED: hypothetical protein n=1 Tax=Vitis vinifera RepID=UPI00019828A8
FRFM4LP02SGPW7	UniRef100_D1H2W6	Whole genome shotgun sequence of line PN40024, scaffold_0.assembly12x (Fragment) n=1 Tax=Vitis vinifera RepID=D1H2W6_VITVI
FRFM4LP02SH0U0	UniRef100_UPI000198489D	PREDICTED: hypothetical protein n=1 Tax=Vitis vinifera RepID=UPI000198489D
FRFM4LP02SIHK1	UniRef100_Q9LKW3	Dehydration-induced protein ERD15 n=1 Tax=Solanum lycopersicum RepID=Q9LKW3_SOLLC
FRFM4LP0	UniRef100_UPI	PREDICTED: hypothetical protein n=1 Tax=Vitis vinifera

2SRQZ7	0001983901	RepID=UPI0001983901
FRFM4LP02SRZML	UniRef100_UPI00019844A0	PREDICTED: hypothetical protein n=1 Tax=Vitis vinifera RepID=UPI00019844A0
FRFM4LP02SSOZK	UniRef100_C0M519	CBL-interacting protein kinase 24 n=1 Tax=Populus euphratica RepID=C0M519_POPEU
FRFM4LP02T0VPI	UniRef100_Q6K7R9	DEAD-box ATP-dependent RNA helicase 48 n=3 Tax=Oryza sativa RepID=RH48_ORYSJ
FRFM4LP02T10OY	UniRef100_D0EWD6	Heat shock protein 90 n=1 Tax=Ipomoea nil RepID=D0EWD6_IPONI
FRFM4LP02T2J7H	UniRef100_UPI000198306B	PREDICTED: hypothetical protein n=1 Tax=Vitis vinifera RepID=UPI000198306B
FRFM4LP02TBVNM	UniRef100_O24363	Phosphorylase n=1 Tax=Spinacia oleracea RepID=O24363_SPIOL
FRFM4LP02TBWD8	UniRef100_Q3HNF4	ABA 8'-hydroxylase CYP707A1 n=1 Tax=Solanum tuberosum RepID=Q3HNF4_SOLTU
FRFM4LP02TCK2T	UniRef100_Q6UK17	HD-ZIP n=1 Tax=Capsicum annuum RepID=Q6UK17_CAPAN
FRFM4LP02TCL3X	UniRef100_A9QVI8	Heat shock protein 90 n=1 Tax=Ageratina adenophora RepID=A9QVI8_9ASTR
FRFM4LP02TEQQS	UniRef100_B9SNH2	Histidine kinase 1, 2, 3 plant, putative n=1 Tax=Ricinus communis RepID=B9SNH2_RICCO
FRFM4LP02TFY70	UniRef100_B9SUH8	Putative uncharacterized protein n=1 Tax=Ricinus communis RepID=B9SUH8_RICCO
FRFM4LP02TH7EP	UniRef100_C5YJ75	Putative uncharacterized protein Sb07g028270 n=1 Tax=Sorghum bicolor RepID=C5YJ75_SORBI
FRFM4LP02THFJL	UniRef100_A8MRX5	Uncharacterized protein At4g19230.2 n=1 Tax=Arabidopsis thaliana RepID=A8MRX5_ARATH
FRFM4LP02TLYIZ	UniRef100_B8YIB0	Homeobox leucine zipper protein n=1 Tax=Mirabilis jalapa RepID=B8YIB0_MIRJA
FRFM4LP02TPE3M	UniRef100_D0EWD6	Heat shock protein 90 n=1 Tax=Ipomoea nil RepID=D0EWD6_IPONI
FRFM4LP02TSJFQ	UniRef100_UPI0001983901	PREDICTED: hypothetical protein n=1 Tax=Vitis vinifera RepID=UPI0001983901
FRFM4LP02TWV1S	UniRef100_C0M519	CBL-interacting protein kinase 24 n=1 Tax=Populus euphratica RepID=C0M519_POPEU
FRFM4LP02TX2FX	UniRef100_Q9AXF5	Starch phosphorylase (Fragment) n=1 Tax=Ipomoea batatas RepID=Q9AXF5_IPOBA
FRFM4LP02TZAXU	UniRef100_UPI000198471B	PREDICTED: hypothetical protein n=1 Tax=Vitis vinifera RepID=UPI000198471B
FRFM4LP02TZFNZ	UniRef100_Q9LKW3	Dehydration-induced protein ERD15 n=1 Tax=Solanum lycopersicum RepID=Q9LKW3_SOLLC